Incentive Regulation and Its Application to Electricity Networks

PAUL L. JOSKOW *

Alfred P. Sloan Foundation and Department of Economics, MIT

Abstract

This paper examines developments since the publication of *The Economics of Regulation* in the theory of incentive regulation and its application to the regulation of unbundled electricity transmission and distribution networks. Conceptual mechanism design issues that arise when regulators are imperfectly informed and there is asymmetric information about costs, managerial effort, and quality of service are discussed. The design and application of price cap mechanisms and related quality of service incentives in the UK are explained. The limited literature that measures the effects of incentive regulation applied to electricity networks is reviewed.

1 Introduction

Alfred Kahn began to write what became *The Economics of Regulation* while I was an undergraduate at Cornell. He was my teacher and academic advisor at Cornell and is the one who stimulated my interest in both economics and the economics of regulation. Much has changed since *The Economics of Regulation* was published in 1970/71 (Volume 1 in 1970 and Volume 2 in 1971). Most of the industries that were thought to be "natural monopolies" and were subject to price, entry and service quality regulation at that time (for example, telecommunication, electricity, natural gas transportation, cable television, etc.) have been restructured and competition introduced into one or more of their horizontal segments.[1] Other industries, where the economic case for pervasive price and entry regulation was already increasingly being recognized as dubious by 1970, and where regulation and competition were often mixed together (for example, trucking, airlines, railroads, natural gas production), have been completely deregulated. The expanse of the economy in the U.S. and most other countries that is subject to price, entry and service quality regulation today has shrunk considerably since 1970.

There are many reasons for this trend, including changes in technology, poor performance exhibited by some regulated industries, changes in the political economy of regulation and associated changes in the power of different interest groups, and broader

* President, Alfred P. Sloan Foundation, 630 Fifth Avenue, Suite 2550, New York, NY. E-mail: joskow@sloan.org This paper is based on a longer study "Incentive Regulation in Theory and Practice: Electric Distribution and Transmission," Joskow (2006a) prepared for the National Bureau of Economic Research regulation project. http://econ-www.mit.edu/files/1181. See also Joskow (2006b). The views expressed here are my own and do not reflect the views of the Alfred P. Sloan Foundation, MIT or any other organization with which I am associated.

[1] Competition in Electricity remains a work in progress in the U.S. See Joskow (2006c).

ideological shifts favoring markets over regulation and state-owned enterprises. While the pendulum may be shifting back in some sectors (for example, financial market regulation, health and safety regulation, access of content providers to communications networks, etc.) the broad changes in the mix of regulated and competitive segments that we have observed in the last 30 years is unlikely to be reversed.

Even in countries that have gone the farthest in "liberalizing" previously regulated and state-owned enterprises, certain network segments of some of the historically regulated "natural monopoly" industries continue to be subject to price, entry and service quality regulation. These industries include electricity transmission and distribution networks, natural gas transmission and distribution networks, and water supply networks. In the case of electricity and natural gas transmission and distribution networks, the regulatory mechanisms applied to these networks have important implications for supporting wholesale competition (electricity generation, natural gas production, and associated wholesale marketing activities) and retail competition (competition to supply end-users) since they serve as platforms upon which competing wholesale and retail suppliers depend and also implement market and non-market mechanisms to maintain network reliability. Thus, good performance of the competitive segments depends on good performance of the remaining regulated network segments.

The application of sound regulatory mechanisms that affect the terms and conditions of network connections, network delivery prices, network investment, and network service quality have been important components of all successful electricity sector liberalization programs around the world. The benefits of a good regulatory framework include lower network service costs, improvements in service quality, investment to expand the network to support changes in supply and demand for network services, and the development of efficient network platforms to support robust competitive wholesale and retail markets. While many of the basic regulatory principles discussed in *The Economics of Regulation* still apply to these remaining regulated monopoly network segments, there have been important advances in both theory and application since those volumes were published as well.

Volume 1 of *The Economics of Regulation* focuses on the principles for pricing regulated services supplied by firms that are subject to budget or break-even constraints. While there have certainly been theoretical advances associated with the second-best pricing of services supplied by regulated monopoly firms since 1970, especially with regard to the design and welfare properties of non-linear prices to meet budget constraints, the application of these basic pricing principles at the retail level has not advanced very far. So, for example, except for large retail customers, time of day pricing and real time pricing for electricity and natural gas has not spread quickly at all, despite the fact that the information available from wholesale markets about generation and natural gas prices makes it even easier to apply these concepts today than in 1970. At the wholesale level, in those places where prices have been deregulated, the market naturally leads to the load varying prices whose basic economic principles are developed in detail in Volume 1.

The problem of designing rewards and incentives for efficient production by firms subject to cost-of-service or rate-of-return regulation is discussed briefly in Volume 1 (pages 53-54) and in (the more rarely read) Volume 2 (Chapters 2 and 3). The discussion in Volume II of *The Economics of Regulation* in particular raises the right issues:

> …the central institutional questions have to do with the nature and adequacy of the incentives and pressures that influence private management in making the critical economic decisions. (Volume II, page 47)

> …[rate of return regulation] creates strong incentives to pad their expenses." (Volume II, page 48)

It also has several insights into how incentives might be introduced into the regulatory process to improve performance:

> Freezing rates for the period of the [regulatory] lag imposes penalties for inefficiency, excessive conservatism, and wrong guesses, and offers rewards for the opposites …

From an overall economic efficiency perspective, these production cost and service quality inefficiencies are likely to be more important than are failures to adopt the most efficient (second best given budget constraints) pricing structures. This is the case because the efficiency losses from excessive costs lead to "first order" efficiency losses ("rectangles") while the pricing inefficiencies are likely to be second order (Harberger "triangles"). The historical focus on efficient price structures, rather than cost control and service quality incentives, likely flows in part from the political concerns about monopoly power, excessive prices, and price discrimination that played an important political role leading to the creation of regulated legal monopolies and oligopolies in the first place. But we must recognize as well that in the last fifteen or twenty years there have been major advances in imperfect and asymmetric information theory, and in the theories of incentive mechanisms and associated contractual arrangements generally that have now made it possible to develop relevant theories and then to apply them.

At the time *The Economics of Regulation* was written, there was relatively little formal theoretical development of the properties of alternative incentive regulation mechanisms that provide incentives to regulated firms to control costs, to offer appropriate levels of service quality, and to find it in their interest to set efficient (second-best) price structures. Absent relevant theory, it was difficult to develop applications that could be applied in the real world, though experiments with incentive regulation go back to the 19th century (Joskow, 2007). At the time *The Economics of Regulation* was published, the primary theoretical analysis that focused on the incentive properties of rate of return regulation was the Averch-Johnson (or as Kahn refers to it in his book, the Averch, Johnson, and Wellisz or A-J-W effect, to recognize the less widely cited paper by Stanislaw Wellisz that identified similar potential distortions from rate of return regulation (Averch and Johnson, 1962; Wellisz, 1963)). The A-J-W effect turns on the incentives created by a characterization of rate of return regulation that effectively reduces the regulated firm's effective cost of capital inputs (r) by creating a profit margin on increases in capital input while leaving fixed the price of other inputs ("labor" in the A-J model) since these input costs are assumed (that is, asymmetrically vis-à-vis capital costs) to be passed through dollar for dollar into regulated prices. This in turn leads a profit maximizing regulated firm subject to this type of regulation to make long run production decisions that use a higher capital/labor ratio than would be cost-minimizing given the firm's production function and true input costs. This theory ignores many attributes of real regulatory institutions and it has little if any empirical support (Joskow, 1974, 2007; Joskow and Rose, 1989), but for many years it was "the" positive theory of regulation. However, in the last fifteen or twenty years there have been significant advances in the theory of "incentive regulation" or "performance-based regulation" and these concepts are beginning to be applied in the regulation of electricity and gas transmission and distribution networks in a number of countries (Joskow 2006a, 2006b, 2007).

The rest of this paper identifies the key elements associated with the development of modern incentive regulation theory and then examines the application of alternative types of "incentive" or "performance-based" regulation of electricity distribution and

transmission network price levels, price structures and service quality. The discussion will assume that effective electricity sector restructuring and unbundling mechanisms have been put in place so that there are clearly defined distribution and transmission network entities offering unbundled delivery and network support services to market participants (as in the UK and portions of the U.S.). I will also assume that electricity networks are regulated monopolies[2] and that an independent regulator with adequate staff resources has been created to oversee the regulation of the distribution and transmission networks.

2 Theoretical considerations

The primary goal of regulation in the public interest is to stimulate the regulated firm to produce output efficiently from cost and quality (including reliability) perspectives, to price the associated services efficiently, and to achieve these goals consistent with satisfying a break-even or budget-balance constraint for the regulated firm that allows the firm to covered its costs of providing service while restraining its ability to exercise its market power to exploit consumers by charging excessive prices. Much of the older theoretical literature on optimal (first and second-best) pricing of services provided by regulated monopolies (for example, Boiteux, Steiner, Turvey) assumes implicitly that regulators are perfectly informed about the regulated firm's cost opportunities and demand patterns and can effectively enforce cost minimization on the regulated firm.[3] The literature then focuses on first and second-best pricing of the services provided by the regulated firm given defined cost functions, demand attributes and budget balance constraints (for example, Ramsey-Boiteux pricing, non-linear pricing, etc.).[4] The older literature did not focus on incentives to minimize costs or improve other dimensions of firm performance (for example, service quality attributes), aside from making the general observation that firms insulated from competition and subject to cost-based regulation were likely to be inefficient and the limited formal theoretical developments of the A-J-W effect discussed above.

In reality, regulators care (or at least should care) as well (or more) about the production efficiency and service quality implications of the regulatory mechanisms they choose. Regulators are neither completely informed nor completely uninformed about relevant cost, quality, and demand attributes faced by the regulated firm. Regulators have *imperfect* information about these firms and market attributes and the regulated firm generally has more information about these attributes than does the regulator. Furthermore, managers have discretion to make choices not only about input proportions (as in the A-J-W models) but on how hard they will work to minimize the firm's costs or in choosing the levels of service quality. Accordingly, the regulated firm may use its information advantage (*asymmetric information)* strategically to exploit the regulatory process to

[2] The economic attributes of unregulated "merchant" transmission network investment are discussed in Joskow and Tirole (2005).

[3] An exception is the extensive theoretical and limited empirical literature following Averch and Johnson (1962), and especially after Baumol and Klevorick (1970) that examines potential distortions in input proportions caused by rate-of-return constraints. The empirical foundations for these theories are discussed in Joskow and Rose (1989).

[4] Brauetigam (1989).

increase its profits or to pursue other managerial goals, to the disadvantage of consumers (Laffont and Tirole, 1993, Chapter 1).

This creates potential moral hazard (for example, too little managerial effort resulting in excessive costs) and adverse selection (for example, prices that are too high relative to production costs) problems that effective regulatory mechanism design must address. The recent theoretical literature on incentive regulation focuses on devising regulatory mechanisms to respond to these moral hazard and adverse selection problems (Laffont and Tirole, 1993; Armstrong and Sappington, 2007).

Consider a situation in which the regulator is uncertain about the firm's true underlying costs and its opportunities further to reduce costs, the regulator cannot observe the level of managerial effort expended by the firm, but the regulator can monitor accurately the firm's realized costs ex post in regulatory hearings and through audits. The regulated firm knows its true cost opportunities, its managerial effort, and the effects of managerial effort on costs. Following Laffont and Tirole (1993, pp.10-19), under these assumptions we can think of two polar case regulatory mechanisms that may be applied to a monopoly firm producing a single product with a fixed quality. The first regulatory mechanism involves setting a fixed price ex ante that the regulated firm will be permitted to charge going forward (that is, effectively forever). In a dynamic setting this is equivalent to a pricing formula that starts with a particular price and then adjusts this price for exogenous changes in input price indices and other exogenous indices of cost drivers (again, effectively forever). This type of regulatory mechanism can be characterized as a fixed price regulatory contract or, in a dynamic setting, a price cap regulatory mechanism.

Because prices are fixed with this mechanism (or vary based only on exogenous indices of cost drivers) and do not respond to changes in managerial effort or ex post cost realizations, the firm and its managers keep 100% of any cost reductions they realize by increasing effort. Accordingly, and ignoring service quality and investment considerations for now, this mechanism provides incentives to induce efficient levels of managerial effort and in turn cost reduction. This effect is a first order "rectangle" efficiency gain. However, because the regulator must ensure that any regulatory mechanism it imposes on the regulated firm meets a budget balance constraint, when the regulator is uncertain about the regulated firm's true cost opportunities she will have to set a relatively high fixed price (or dynamic price cap) to ensure that *if* the firm is indeed inherently high cost, the prices under the fixed price contract or price cap will be high enough to cover the firm's (efficient but high) realized costs. Accordingly, while a fixed price mechanism does well from the perspective of providing incentives to reduce costs it is potentially very poor at "rent extraction" for the benefit of consumers and society because prices may be too high relative to the firm's true cost opportunities. The social value of rent extraction depends upon the social welfare function applied to the distribution of these rents between consumers and producers (Armstrong and Sappington, 2007) or the cost of public funds in a public procurement theoretical framework (Laffont and Tirole, 1993).

At the other extreme, the regulator could implement a simplistic pure "cost of service" regulatory contract where the firm is assured that it will be compensated for all of the costs of production that it actually incurs and no more. After the firm produces, the regulator's uncertainty about whether the firm is a relatively high or a low cost opportunity firm will be resolved. And since the regulator compensates the firm only for its realized costs, there is no "rent" left to the firm or its managers in the form of excess profits. This solves the "rent extraction" or "adverse selection" problem that would arise under a fixed price

contract. However, this kind of cost of service regulatory mechanism does not provide any incentives for the management to exert optimal (indeed any) effort. Even though there are no "excess profits" left to the firm, the actual costs incurred by the firm may be inefficiently high as a result of too little managerial effort. Managers now retain 0% of any cost savings they achieve and have no incentive to exert cost-reducing effort. Accordingly, consumers may now be paying higher prices than they would have to pay if the management could be induced to exert more effort to reduce costs. Indeed, it is this kind of managerial slack and associated x-inefficiencies that most policymakers have in mind when they discuss the "inefficiencies" associated with regulated firms.

Conceptually, fixed-price contracts (or price caps) are good at providing incentives for managerial efficiency and cost minimization, but bad at extracting the benefits of the lower costs for consumers. Cost of service contracts are good at aligning prices and costs but the costs will be excessive due to suboptimal managerial effort. Perhaps not surprisingly, the optimal regulatory mechanism in the presence of imperfect and asymmetric information will lie somewhere between these two extremes. It will have a form similar to a *profit sharing* contract or a *sliding scale* regulatory mechanism where the price that the regulated firm can charge is *partially* responsive to or contingent on changes in realized costs and *partially* fixed ex ante (Schmalensee, 1989; Lyon, 1996). (I should note that Volume II of *The Economics of Regulation* discusses some early profit sharing or sliding scale plans and performance benchmarking mechanisms (pp.61-63), but expresses some skepticism about the regulator's ability to apply these mechanisms effectively.) More generally, by offering the regulated firm a *menu* of cost-contingent regulatory contracts with different cost sharing provisions, the regulator can do even better than if it offers only a single profit sharing contract (Laffont and Tirole, 1993).

3 Price cap mechanisms in practice

While the theoretical literature on incentive regulation is quite rich, it still provides relatively little direct guidance for practical application in real-world circumstances. In practice, well-designed incentive regulation programs have adopted fairly simple mechanisms that reflect some of the basic theoretical principles discussed above.

A particular form of incentive regulation was introduced for the regulated segments of the privatized electric gas, telephone and water utilities in the UK, New Zealand, Australia, and portions of Latin American as well as in the regulated segments of the telecommunications industry in the U.S.[5] This mechanism chosen is the "price cap" (Beesley and Littlechild, 1989; Brennan, 1989; Armstrong, Cowan and Vickers, 1994; Isaac, 1991). Price cap regulation is a form of institutionalized regulatory lag. Under price cap regulation the regulator sets an initial price p_o (or a vector of prices for multiple products). This price (or a weighted average of the prices allowed for firms supplying multiple products or different types of customers) is then adjusted from one year to the next for changes in inflation (rate of input price increase or RPI) and a target productivity change factor "x."[6] Accordingly, the price p_1 in period 1 is given by:

[5] The U.S. is behind many other countries in the application of incentive regulation principles to electric distribution and transmission, though their use is slowly spreading in the U.S. beyond telecommunications.
[6] Many implementations of price cap regulation also have "z" factors. Z factors reflect cost elements that cannot be controlled by the regulated firm and are passed through in retail prices. For example, in the UK,

(1) $$p_1 = p_o (1+ RPI – x)$$

In theory, a "forever" price cap mechanism is a high-powered "fixed price" regulatory contract which provides powerful incentives for the firm to reduce costs. Moreover, if the price cap mechanism is applied to a (properly) weighted average of the revenues the firm earns from each product it supplies, the firm has an incentive to set the second-best prices for each service (Laffont and Tirole, 2000) given the level of the price cap. So to speak, it kills two birds with one stone. As already noted, however, when the regulator has imperfect information about the firm's cost opportunities and must meet a budget balance constraint, pure "forever" price cap mechanisms are not optimal from the perspective of an appropriate tradeoff between efficiency incentives and rent extraction (Schmalensee, 1989) and would leave too much rent to the firm with "average" cost characteristics. Finally, any incentive regulation mechanism that provides incentives only for cost reduction also potentially creates incentives inefficiently to reduce service quality when service quality and costs are positively correlated with one another.

In practice, "forever" price caps are not typically used in the regulation of distribution and transmission network price levels. Some form of cost-based regulation is used to set an initial value for p_o. The price cap mechanism then operates for a pre-established time period (for example, five years). At the end of this period a new starting price p_o and a new x factor are established after another cost-of-service and prudence or efficiency review of the firm's costs. That is, there is a pre-scheduled regulatory process to reset or "ratchet" prices based partially on costs realized during the previous period. In addition, price caps are often only one component of a larger portfolio of incentive mechanisms that include quality of service incentives, as discussed in the next section. Finally, regulated electric distribution and transmission network firms' ability to determine the structure of prices for different types of customers or for services provided at different locations on the network under an overall revenue cap is typically limited. As a result, the applications of price caps in practice are properly thought of as cost and quality incentive mechanism not as a mechanism to induce optimal second-best pricing of various network services. So, in practice the incentive mechanisms are only targeted at one bird rather than two.

A natural question to ask about price cap mechanisms is where does "x' (and perhaps p_o) come from or, more generally, how does one choose the correct starting value for p_o and the proper dynamic price trajectory? The difficulty of answering this question in practice is one of the sources of skepticism about formal incentive mechanisms expressed in Volume II of *The Economics of Regulation.* In England and Wales and some other countries, statistical benchmarking methods have come to be used to help to determine the relative efficiency of individual firms' operating costs and service quality compared to their peers. This information can then be used as an input to setting values for both p_o and x (Jamasb and Pollitt, 2001, 2003; OFGEM, 2004a) to provide incentives for those far from the efficiency frontier to move toward it and to reward the most efficient firms in order to induce them to stay on the efficiency frontier, in a fashion that is effectively an application of yardstick regulation (Shleifer, 1985).

Although it is not discussed too much in the theoretical or empirical literature on price caps, capital-related cost are handled quite differently from operating costs in the

---

the charges distribution companies pay for connections to the transmission network are treated as pass-throughs. Changes in property tax rates are also often treated as pass-throughs.

establishment and resetting of p_o and x. The limited attention paid to capital-related costs in the academic literature on price cap regulation provides a potentially misleading picture of the challenges associated with implementing a price-cap mechanism effectively. This is the case for several reasons. First, in practice, the p_o and x values must be developed based not only on a review of the relative efficiency of each firm's operating costs, but also based on the value of the firm's current capital stock or rate base, forecasts of future capital additions required to provide target levels of service quality, and the application of depreciation rates, estimates of the cost of the firm's debt and equity capital, assumptions about the firm's dept/equity ratio, tax allowances and other variables to turn capital stocks into prices for capital services over time. The capital cost related allowances represent a large fraction of the total price (p_o) of supplying unbundled electricity network services so the choices of these parameters for defining capital user charges are very important.

Second, allowances for capital-related costs are typically established by regulators using incentive regulation mechanisms through more traditional utility planning and cost-of-service regulatory accounting methods including the specification of a rate base (or regulatory asset value), depreciation rates, debt and equity costs, debt/equity ratios, tax allowances, etc. This is the case because the kinds of statistical benchmarking techniques that have been applied to operating costs have not been developed for capital-related costs, due to significant heterogeneity between firms in terms of the age of assets, geography, service quality, lumpiness of capital investments and other considerations. Third, the efficiency properties of a regulatory mechanism that mixes competitive benchmarking with more traditional forward-looking rate of return regulation are more complex than first meets the eye (Acemoglu and Finkelstein, 2006).

In principle, operating and capital costs could be integrated and associated benchmarks determined using total factor productivity measures. This is the approach taken by the initial price caps applied to telecom companies in the U.S. by the Federal Communications Commission. In electricity, this approach has been rejected largely because of the diversity in the capital stock, much of which is several decades old, and the associated difficulties of coming up with accurate total productivity measures. The application of price caps in England and Wales and other countries in Europe that have adopted this mechanism, benchmark a firm's performance against industry specific "best practice" (production frontier analysis using data for other firms in the industry).

Thus, the implementation of price cap mechanisms is more complicated and their efficiency properties more difficult to evaluate than is often implied and places a significant information collection, auditing and analysis burden on regulators. This is precisely the source of the skepticism about formal incentive mechanisms expressed in Volume II of *The Economics of Regulation*. In practice, modern applications of incentive regulation concepts involve the application of elements of traditional cost of service regulation, yardstick regulation, and high-powered "fixed price" incentives.

The challenge of forecasting future investment needs and costs for electricity network firms has historically been a rather contentious process, sometimes yielding significant differences between what the regulated firm's claim they need and what the regulator claims they need to meet their legal responsibilities to provide safe and reliable service efficiently. There is clearly a very serious asymmetric information problem here. In the 2004 review of electricity distribution prices in the UK, the regulator adopted an innovative "menu" of sliding scale mechanisms approach to resolve the asymmetric information problem faced by the regulator as she tries to deal with differences between

the firms' claims and the regulator's consultants' claims (OFGEM, 2004b) about future capital investment requirements to meet reliability targets. The sliding scale menu allows firms to choose between getting a lower capital expenditure allowance but a higher powered incentive (and a higher expected return on investment) that allows them to retain more of the cost reduction if they can beat the target expenditure levels or a higher capital expenditure allowance combined with a lower powered sliding scale mechanism and lower expected return. (OFGEM, 2004b). This is an application of Laffont and Tirole's menu of cost-contingent contracts mechanism and provides a more effective way to deal with the imperfect and asymmetric information conditions and associated adverse selection problems than the traditional approach of offering a single regulatory contract.

An example of the use of a profit-sharing or cost-contingent form of incentive regulatory mechanism can be found in the incentive mechanism that has been applied to the costs of the transmission system operator (SO) in England and Wales (which is also the transmission owner (TO), though there are separate regulatory mechanisms for SO and TO functions). Each year forward targets are established for the costs of system balancing services and system losses (OFGEM, 2005). A sharing or sliding scale formula is specified which places the TO at risk for a fraction (for example 30%) of deviations from this benchmark (up or down) with caps on profits and losses. There is also a cap and a floor. In recent years the SO was given a menu of three alternative incentive arrangements with different sharing fractions and different caps and floors (with costs of service as a default) from which to choose. If the SO were to choose the cost-of-service default it would suggest that in constructing the menu, the regulator had underestimated the range of the SO's future cost realizations.

4 Service quality incentives

As noted earlier, any incentive regulation mechanism that provides incentives only for cost reduction also potentially creates incentives to reduce service quality when service quality and costs are positively related to one another. The higher powered are the incentives to reduce costs, the greater the incentive to reduce quality when cost and quality are correlated. Accordingly, price cap mechanisms are increasingly accompanied by a set performance standards and associated penalties and rewards for the regulated firm for falling above or below these performance norms. Similar mechanisms are used by several U.S. states and in other countries that have liberalized their electricity sectors (for example, New Zealand, Netherlands, and Argentina).

In the UK, the regulator (OFGEM) has developed several incentive mechanisms targeted at various dimensions of distribution network service quality (OFGEM, 2004b, 2004c). OFGEM uses statistical and engineering benchmarking studies and forecasts of planned maintenance outages to develop targets for the number of customer outages and the average number of minutes per outage for each distribution company.

Until recently in the UK, there was no formal incentive mechanism that applied to transmission system reliability – network failures that lead to administrative customer outages or "unsupplied energy". In 2005, a new incentive mechanism that focuses on the reliability of the transmission network as measured by the quantity of "unsupplied energy" resulting from transmission network outages went into effect (OFGEM, 2004d). NGC is assessed penalties or received rewards when outages fall outside of a "deadband" of +/- 5%

defined by the distribution of historical outage experience (and with potential adjustments for extreme weather events), using a sliding scale with a cap and a floor on the revenue impact.

5 Performance attributes

The information burden to implement incentive regulation mechanisms well is certainly no less than for traditional cost of service regulation. Incentive regulation in practice requires a good accounting system for capital and operating costs, cost reporting protocols, data collection and reporting requirements for dimensions of performance other than costs. Capital cost accounting rules are necessary, a rate base for capital must still be defined, depreciation rates specified, and an allowed rate of return on capital determined. Comprehensive "rate cases" or "price reviews" are still required to implement "simple" price cap mechanisms. Planning processes for determining needed capital additions are an important part of the process of setting total allowed revenues and associated prices going forward. Performance benchmarks must be defined and the power of the relevant incentive mechanisms determined. What distinguishes incentive regulation in practice from traditional cost of service regulation is that this information is used more effectively because it can rely on advances in incentive regulation theory to organize and apply it. Whether the extra effort is worth it depends on whether the performance improvements justify the additional effort.

Unfortunately, there are been relatively little systematic analysis of the effects of the application of incentive regulation mechanisms on the performance of electric distribution and transmission companies.[7] Improvements in labor productivity and service quality have been documented for electric distribution systems in England and Wales, Argentina, Chile, Brazil, Peru, New Zealand and other countries (Newbery and Pollitt, 1997; Rudnick and Zolezzi, 2001; Bacon and Besant-Jones, 2001; Estache and Rodriguez-Pardina, 1998; and Pollitt, 2004). However, most of these studies have focused on developing countries where the pre-reform levels of performance were especially poor prior to restructuring. Moreover, it is difficult to disentangle the effects of privatization, restructuring and incentive regulation from one another.

The most comprehensive study of the post reform performance of the regional electricity distribution companies in the UK (distribution and supply functions) has been done by Domah and Pollitt (2001). They find significant overall increases in productivity over the period 1990 to 2000 and lower real "controllable" distribution costs compared to a number of benchmarks. However, controllable costs and overall prices first rose in the early years of the reforms before falling dramatically after 1995. Moreover, the first application of price cap mechanisms to the distribution networks in 1990 was too generous (average of RPI+ 2.5%) and a lot of rent was initially left on the table for the RECs' initial owners (who cleverly soon sold out to foreign buyers). Distribution service quality in the UK, at least as measured by supply interruptions per 100 customers and average minutes of service lost per customer, has improved as well since the restructuring and privatization initiative in 1990. This suggests that incentive regulation has not led, as some had feared,

[7] There is a much more extensive body of empirical work that examines the effects of incentive regulation mechanisms, primarily price caps, on the performance of telecommunications firms. Examples are Ai and Sappington (2004), Sappington (2003), Ai, Martinez and Sappington (2005).

to deterioration in these dimensions of service quality. This is likely to have been the case because quality standards and associated mechanisms were included in the portfolio of incentive regulation mechanisms adopted in the UK.

The experience with the transmission system operator (SO) incentive mechanism in England and Wales also provides a good example of how incentive regulation can improve performance. During the first few years following the restructuring of the electricity sector in England and Wales in 1990, the SO recovered the costs of system balancing, including managing congestion and other network constraints, through a simple cost pass-through mechanism. The SO's costs escalated rapidly, growing from about $75 million per year in 1990/91 to almost $400 million per year in 1993/94. After the introduction of the SO incentive scheme in 1994, these costs fell to about $25 million in 1999/2000. OFGEM estimates that NGC's system operating costs fell by about £400 million ($600 million at current exchange rates) between 1994 and 2001. A new SO incentive scheme was introduced when NETA went into operation in early 2001. The SO's costs have fallen by nearly 20% over the three year period since the new scheme was introduced (OFGEM, 2003).

While the empirical evidence on the effects incentive regulation mechanisms applied to electric distribution and transmission system in practice is still limited, the experience so far is very encouraging.

6 Conclusion

As we look back at developments in the theory and practice of regulating firms that have been given de facto monopoly franchises since the publication of *The Economics of Regulation* in 1970 and 1971, we can come to a number of conclusions. First, the overall economic importance of getting the theory and application of "natural monopoly" regulation right has become less important over time since a smaller fraction of the economy is subject to these types of economic regulation as competition has replaced regulated monopoly in so many industries. Moreover, even in those industries where price, entry and service quality regulation continues, it is typically being applied to a smaller number of truly natural monopoly network segments of those industries as competition has been introduced into other horizontal segments that rely on the regulated network platform. The basic theory of efficient pricing (first and second best) under the assumption that the regulator is fully informed has not changed very much over the years, except perhaps for a better understanding of the properties of non-linear pricing for regulated firms subject to budget constraints. The major advances in the theory and practice of regulation have relied on formalizing the information structure that characterizes the real world. Regulators are imperfectly informed, regulated firms have better information about the cost and demand attributes they face, and regulated firms will use this information advantage to their benefit. This situation leads to adverse selection and moral hazard problems that have been incorporated into the modern theory of incentive regulation. While applying this theory in practice must confront numerous empirical challenges, the available evidence from their application to electricity distribution and transmission systems suggests that they can help to resolve what Kahn called "the central institutional question" that confronts economic regulation.

7 References

Acemoglu, D. and A. Finkelstein (2006) "Input Technology Choice in Regulated Industries: Evidence from the Health Care Sector," mimeo, MIT, Cambridge, February.

Ai, C. and D. Sappington (2005) "Reviewing the Impact of Incentive Regulation on U.S. Telephone Service Quality," *Utilities Policy*, 13: 201-210.

Ai, C., Martinez, S., and D. Sappington (2004) "Incentive Regulation and Telecommunications Service Quality," *The Journal of Regulatory Economics*, 26:263-285.

Armstrong, M., S. Cowan and J. Vickers (1994) *Regulatory Reform: Economic Analysis and British Experience*, MIT Press: Cambridge, MA.

Armstrong, M. and D. Sappington (2007) "Recent Developments in the Theory of Regulation," in M. Armstrong and R. Porter (ed.), *Handbook of Industrial Organization (Vol. III)*, Elsevier Science Publishers: Amsterdam.

Averch, H. and L.L. Johnson (1962) "Behavior of the Firm under Regulatory Constraint," *American Economic Review*, 52: 1059-69.

Bacon, R. W., and J. E. Besant-Jones (2001) "Global Electric Power Reform, Privatization and Liberalization of the Electric Power Industry in Developing Countries," *Annual Reviews of Energy and the Environment*, 26: 331–59.

Baumol, W. and A.K. Klevorick (1970) "Input Choices and Rate of Return Regulation: An Overview of the Discussion," *Bell Journal of Economics and Management Science*, 1: 169-190.

Beesley, M. and S. Littlechild (1989) "The Regulation of Privatized Monopolies in the United Kingdom," *Rand Journal of Economics*, 20: 454-472.

Braeutigam, R. (1989) "Optimal Prices for Natural Monopolies," in R. Schmalensee and R. Willig (ed.), *Handbook of Industrial Organization, Volume II*, Elsevier Science Publishers: Amsterdam.

Brennan, T. (1989) "Regulating By Capping Prices," *Journal of Regulatory Economics*, 1: 133-147.

Domah, P.D. and M.G. Pollitt (2001) "The Restructuring and Privatisation of the Regional Electricity Companies in England and Wales: A Social Cost Benefit Analysis," *Fiscal Studies*, 22:107-146.

Estache, A. and M. Rodriguez-Pardina (1998) "Light and Lightening at the End of the Public Tunnel: The Reform of the Electricity Sector in the Southern Cone," World Bank Working Paper, May.

Isaac, R.M. (1991) "Price Cap Regulation: A Case Study of Some Pitfalls of Implementation," *Journal of Regulatory Economics*, 3: 193-210.

Jasmsb, T. and M. Pollitt (2001) "Benchmarking and Regulation: International Electricity Experience, *Utilities Policy*, 9: 107-130.

Jamasb, T. and M. Pollitt (2003) "International Benchmarking and Regulation: An Application to European Electricity Distribution Utilities," *Energy Policy*, 31: 1609-1622.

Joskow, P.L. (1974) "Inflation and Environmental Concern: Change in the Process of Public Utility Price Regulation," *Journal of Law and Economics*, 17: 291-327.

Joskow, P.L. (2006a) "Incentive Regulation in Theory and Practice: Electric Distribution and Transmission," prepared for the National Bureau of Economic Research regulation project. http://econ-www.mit.edu/faculty/download_pdf.php?id=1220

Joskow, P.L. (2006b) "Incentive Regulation for Electricity Networks," *CESifo Dice Report*, 2/2006.

Joskow, P.L. (2006c) "Markets for Power in the United States: An Interim Assessment," *The Energy Journal*, 27: 1-36.

Joskow, P.L. (2007) "Regulation of Natural Monopolies," in A.M. Polinsky and S. Shavell (ed.), *Handbook of Law and Economics (Volume 2)*, Elsevier: Amsterdam.

Joskow, P.L. and N.L. Rose (1989) "The Effects of Economic Regulation," *Handbook of Industrial Organization (Volume II)*, R. Schmalensee and R Willig editors, North-Holland, Amsterdam.

Joskow, P.L. and J. Tirole (2005) "Merchant Transmission Investment," *Journal of Industrial Economics*, 53: 233-64.

Kahn, A.E. (1970, 1971) *The Economics of Regulation: Principles and Institutions*, Volume 1 (1970) and Volume 2 (1971), Wiley: NewYork.

Laffont, J-J and J. Tirole (1993) *A Theory of Incentives in Regulation and Procurement.* MIT Press: Cambridge, MA.

Laffont, J-J and J. Tirole (2000) *Competition in Telecommunication*. MIT Press: Cambridge, MA.

Lyon, T., (1996) "A Model of the Sliding Scale," *Journal of Regulatory Economics*, 9: 227-247.

Newbery, D., and M. Pollitt, (1997) "The Restructuring and Privatization of Britain's CEGB – Was it Worth It?" *Journal of Industrial Economics*, 45: 269-303.

Office of Gas and Electricity Markets (OFGEM) (2003) "NGC System Operator Incentive Scheme from April 2004, Initial Consultation Document," December 2003. London.

Office of Gas and Electricity Markets (OFGEM) (2004a) "Electricity Distribution Price Control Review," Initial Proposals, June 2004, 145/04. London.

Office of Gas and Electricity Markets (OFGEM) (2004b) "Electricity Distribution Price Control Review: Final Proposals," 265/04, November, London.

Office of Gas and Electricity Markets (OFGEM) (2004c) "Electricity Distribution Price Control Review: Appendix – The Losses Incentive and Quality of Service, June 2004 145e/04. London.

Office of Gas and Electricity Markets (OFGEM) (2004d) "Electricity Transmission Network Reliability Incentive Scheme: Final Proposals," December, London.

Office of Gas and Electricity Markets (OFGEM) (2005) "NGC System Operator Incentive Scheme from April 2005," Final Proposals and Statutory License Consultation, March 2005 65/05. London.

Pollitt, Michael, (2004) "Electricity Reforms in Argentina: Lessons for Developing Countries," CMI Working Paper 52, Cambridge Working Papers in Economics. http://www.econ.cam.ac.uk/electricity/publications/wp/ep52.pdf

Rudnick, H., and J. Zolezzi. (2001) "Electric Sector Deregulation and Restructuring in Latin America: Lessons to be Learnt and Possible Ways Forward," in *IEEE Proceedings Generation, Transmission and Distribution*, 148: 180-84.

Sappington, D. (2003) "The Effects of Incentive Regulation on Retail Telephone Service Quality in the United States," *Review of Network Economics*, 2: 355-375.

Schmalensee, R. (1989) "Good Regulatory Regimes," *Rand Journal of Economics*, 20: 417-436.

Schleifer, Andrei (1985) "A Theory of Yardstick Competition," *Rand Journal of Economics*, 16: 319-327.

Wellisz, S. (1963) "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *Journal of Political Economy*, 71: 30-43.

Alfred Kahn As A Case Study of A Political Entrepreneur:

An Essay in Honor of His 90th Birthday

PHILIP J. WEISER *

Professor of Law and Telecommunications, University of Colorado

Abstract

The success of airline deregulation challenged the claims of public choice theory, which asserts that regulation serves the purposes of the regulated firms themselves. One prominent explanation for airline deregulation is that "political entrepreneurs" can, under certain circumstances, challenge the status quo. Nonetheless, commentators have failed to examine how Fred Kahn fits the model of a political entrepreneur. By examining Kahn's success as a political entrepreneur, this Essay highlights how commentators and policymakers can gain important insights into how to spur regulatory reform and oversee regulatory reform efforts like that necessary to modernize our system of spectrum management.

1 Introduction

On many accounts, Alfred Kahn is the seminal figure in regulation in the United States during the late 20th and early-21st century (McCraw, 1984). As a regulator, as an academic, and as an advocate, Fred Kahn has been preeminent in each of these roles, and shaped the way that institutions behave, academics study and, most importantly, regulators regulate. Not only did he write the book (or a two-volume treatise, as it were) on the *Economics of Regulation* (Kahn, 1970), he also wrote and executed the play book on how to conceive of and manage a major deregulatory initiative as a "political entrepreneur." In so doing, he underscored that ideas and effective public administration can make a difference—proving wrong those skeptics who doubted that regulatory reform could ever be accomplished over the wishes of vested interests who benefit from the status quo.[1]

Kahn's leadership as a political entrepreneur is a topic that commentators generally gloss over, offering only a sentence or two to note that Kahn fits the prototype (Peltzman et al., 1989). The concept of a political entrepreneur continues to be developed and is currently used to cover a range of actors, including consumer advocates like Ralph Nader (who famously pushed for auto safety regulations), politicians like Ed Muskie (who spearheaded the drive for environmental protection), and, in a few cases, regulators like

* Contact Author. University of Colorado Law School. 404 Wolf Law Building, 401 UCB, Boulder, CO 80309-0401. E-mail: phil.weiser@colorado.edu

[1] Notably, I do not take a position on the costs or benefits of airline deregulation, but rather, I highlight that the very success of the initiative is worth studying as a model for effectuating regulatory reform.

Fred Kahn (Pozen, 2008).[2] In most cases, discussions of political entrepreneurship tend to be more abstract and not grounded in the practical experience of policymakers like Fred Kahn.

This Essay emphasizes the importance of examining case studies in political entrepreneurship. In particular, as to Fred Kahn, a case study of his leadership underscores the importance of developing a public case for regulatory reform, stretching the legal authority of a regulatory agency to pursue reform objectives, and seeking to empower entities that would benefit from reform. Although this Essay does not seek to undertake an exhaustive study of how Kahn embodied these qualities in his leadership of the Civil Aeronautics Board (CAB), Part 2 explains the political entrepreneur concept and highlights how Kahn's leadership at the CAB provides an exemplary case of political entrepreneurship. Building on this explanation, Part 3 suggests some relevant strategies for political entrepreneurship in the area of spectrum policy reform. Finally, Part 4 offers a short conclusion, calling for a greater investigation of Kahn's model of political entrepreneurship and an increased focus on the role of such leadership in advancing the cause of regulatory reform.

2 The Path of Airline Deregulation and Kahn's Political Entrepreneurship

The concept of a political entrepreneur takes, as a starting point, the premise that individual and collective initiative can influence policy outcomes. On more economically determined views of history, such individual initiative is irrelevant. For strict adherents to public choice theory, for example, it is an article of faith that powerful rent-seeking forces use regulation to protect concentrated interests over diffuse ones, thereby making swimming against that tide both foolhardy and irrational. But such a public choice perspective, as the last thirty years of deregulation have shown, is over-determined and needs to be tempered by, among other things, the political entrepreneur concept.

Scholars are still developing the concept of "political entrepreneurship." Over the last three decades, scholars have discussed the concept or variants of it (such as "policy entrepreneurs"), but it remains an under-developed concept.[3] On many accounts, James Q. Wilson (1975) often receives the credit for the concept, developing it as part of a four part typology (Wilson, 1980).[4] Notably, Wilson recognized the possibility of "entrepreneurial politics" whereby a political entrepreneur, such as Ralph Nader, can overcome public choice forces and mobilize diffuse interests to support regulation over the opposition of

[2] For an even broader conception of the term, see Kingdon (1995, p.204), where he explains that "[political] entrepreneurs are found at many locations; they might be elected officials, career civil servants, lobbyists, academics, or journalists. No one type of participant dominates the pool of entrepreneurs."

[3] As Pozen (2008, p.301) put it, "the phrase 'policy entrepreneur' is [often] invoked without any explanation or references, presumably on the belief that the concept is either self-explanatory or so well known as to make citation pedantic. Not infrequently, authors will use 'policy' interchangeably with 'public,' 'political,' 'bureaucratic,' 'administrative,' or some other variant." Michael Mintrom (2000, p.ix) related, in a similar vein, that although they "have been mentioned in the literature on policymaking, policy entrepreneurs remain ghost-like figures, in need of clearer definition."

[4] Peter Schuck (1981) credits Wilson with developing the concept, but others developed it around the same time (see Walker, 1974).

concentrated interests (i.e., the regulated sector).[5] In short, a quintessential political entrepreneur, like Fred Kahn at the CAB, demonstrates that ideas matter in by politics by implementing an effective political strategy to spur policy reform.[6]

For my purposes, there are three essential elements—all embodied by Kahn's leadership at the CAB—that define the role of a political entrepreneur in spurring policy reform. As a starting point, the critical strategic goal of any political entrepreneur is to develop a sufficient level of public awareness of and support for a policy objective. In Kahn's case, his ability to make the case for reform publicly and with intellectual credibility was essential to his success—not unlike the ability of an actual inventor to make the case for commercializing her invention. In tandem with the ability to make the case for deregulation to the public, Kahn also pursued the objective of eroding the airline industry's commitment to the legacy regulatory regime by both undermining the manner in which it protected established incumbents and bolstering the strength of those interests that would benefit from deregulation. Roughly analogous to the role of market entrepreneurs, the challenge of articulating the case for reform parallels the need to make the case for funding a business venture whereas the second two objectives present execution challenges that, if managed effectively, demonstrate that the venture can and should succeed.

Kahn's success in pursuing these three goals resulted from a unique set of skills and abilities that he brought to the effort. First, he understood and was an evangelist for airline deregulation. Like a private sector entrepreneur, a political entrepreneur need not be the inventor of a policy strategy, but it is essential that she fully understands it and can credibly explain it. Second, Kahn's level of commitment was singular and determined. In this respect, he understood and embodied Max Weber's (1946, p. 128) description of politics as the "strong and slow boring of hard boards. It takes both passion and perspective."

In 1977, when Kahn was appointed Chairman of the Civil Aeronautics Board (CAB), the incumbent airlines constituted a cartel that the CAB, in effect, protected by insisting on uniform pricing (thereby preventing firms from detecting and engaging in price-cutting) and preventing new entry (thereby preventing them from maverick firms). By pursuing a deregulatory initiative against the wishes of the established firms, Kahn challenged the premises of public choice theory, which predicted that any such initiative was doomed to fail. According to public choice theory and, in particular, the allied concept of capture theory, concentrated interests (like the established airline companies) benefit from and dictate the course of regulation to protect their economic rents at the expense of diffuse interests (namely, average consumers).[7]

---

[5] Wilson's (1980) typology provides for three other categories of politics. On his view, public choice politics can be divided into two parts—interest group politics, whereby narrowly dispersed costs and benefits lead to bargaining among affected interests; and client politics, whereby the benefits are concentrated and costs are dispersed. Wilson also highlights that some laws, such as the Social Security Act and the Sherman Act, apply across a wide array of affected parties and are the result of majoritarian politics.

[6] Other commentators have offered a number of different definitions of the concept. Russell Hardin (1982, p.35), for example, explains that "political entrepreneurs are people who, for their own career reasons, find it in their private interest to work to provide collective benefits to relevant groups." Sam Peltzman (1989, p.51), by contrast, focuses on the function of political entrepreneurs, defining them as persons who discover "how to take advantage of the fundamental instability of majority rule within the constraints imposed by the institutional arrangements designed to induce stability."

[7] See Levine (2006, p.270), who related that "airline deregulation was a policy that principally benefited poorly organized consumers and was adopted over the opposition of a relatively small and very well organized group of regulated airlines."

By late 1970s, public choice theory had discredited the public interest theory of regulation. Its popularity undoubtedly reflected the views among policy observers that it accurately described the ways in which many regulatory agencies operated (Schuck, 1981). Reflecting this perspective, Migue (1977, p.213) suggested that "[i]t seems fair to say that among economists the most widely accepted theory of government regulation is that, as a rule, regulation is acquired by the industry regulated and is designed and operated primarily for its benefit." George Stigler (1971, p 5), who was a pioneer in public choice theory, stated similarly that "regulation is acquired by the industry and is designed and operated primarily for its benefit." By successfully overseeing the deregulation of the airlines, Kahn's leadership challenged the premises and predictive power of this theory.[8] Indeed, as Peter Schuck (1981) put it, the success of airline deregulation undermined the stronger versions of public choice theory such that no refinements could rehabilitate it.

In recounting the events that led to the deregulation of the airline industry, it is clear that a number of forces and actors came together to spur the enactment of the Airline Deregulation Act of 1978 (McCraw, 1984). On all accounts, however, Fred Kahn plays a starring role (McCraw, 1984). For future would-be political entrepreneurs, Kahn's leadership strategies and tactics thus bear close analysis.

The appointment of a world class economist to the CAB defied the tradition of appointing individuals from political backgrounds to regulatory agencies.[9] In the wake of Watergate, President Carter took seriously his commitment to restore confidence in government and thus broke from precedent by seeking out individuals with intellectual fortitude, expertise, and independence from traditional regulatory politics.[10] Fred Kahn fit this bill perfectly—with acclaimed experience as a (if not the) leading economist of regulation and as a respected regulator at the New York State Public Service Commission—and he approached his role at the Civil Aeronautics Board with the spirit of a true reformer. Moreover, as an outsider to the airline industry, he viewed both the traditional regulatory agenda and the traditional regulatory processes with skepticism and suspicion. In so doing, he astonished airline executives by making no effort to learn the details of their industry, famously remarking that "I really don't know one plane from the other. To me they are all marginal costs with wings (McCraw, 1984, p.224)."

As to Kahn's substantive agenda, he sought to undermine the traditional premises of airline regulation that competition was destructive and should be avoided through command and control regulation. Rather, he sought to act based on the economic analysis he set forth in his masterful *Economics of Regulation* (Kahn, 1970).[11] In the second

[8] An early contribution to public choice theory came in Mancur Olson, The Logic Of Collective Action: Public Goods And The Theory Of Groups 1-4 (Harvard University Press 1965). In the early 1970s, a number of seminal articles developed the concept. See George J. Stigler, The Theory of Economic Regulation, 2 Bell J. Econ. & Mgmt. Sci. 3 (1971); see also Richard A. Posner, Taxation By Regulation, 2 Bell J. Econ. & Mgmt. Sci. 22 (1971); see also Richard A. Posner, Theories of Economic Regulation, 5 Bell J. Econ. & Mgmt Sci. 335 (1974).

[9] See J. Landis, Report on Regulatory Agencies to the President-Elect (1960) for a description of the type of individuals appointed to regulatory agencies.

[10] When appointing others to the agency, Carter followed Kahn's guidance to avoid appointing a politician and to select individuals with "a firm, independent professional background and qualifications, and is in a position to make strong, intellectually-grounded professional judgments." (McCraw, 1984, p.289)

[11] See also Michael E. Levine, Airline Competition in Deregulated Markets: Theory, Firm Strategy, and Public Policy, 4 Yale J. Reg. 393, 394 (1987) for a description of how "by the mid-1970's it was probably fair to say that no impartial academic observer of any standing doubted that the airline business, if unregulated, would reach something that more or less resembled a competitive equilibrium."

volume of that treatise, Kahn (1971) cited experiments in airline competition undertaken at the state level that underscored the opportunity for discount airline operators to attract new customers and fly planes at much higher load factors than the industry average.

Immediately upon arriving at the CAB, Kahn sought to revolutionize the ways that the agency operated. For starters, he required all communication and orders to be written in understandable prose. In terms of process, Kahn criticized what he viewed as an intellectually bankrupt means of doing business—deciding issues in secret, without deliberation, and asking lawyers to develop the necessary justification for a pre-determined result.[12] In terms of agency structure, he also reorganized its operations, increased the role of professional economists, and appointed very talented deregulatory minded officials. Like at the New York State Public Service Commission, his leadership at the CAB was not hierarchical, involved recruiting and empowering talented staff, and included efforts to persuade existing professional staff of the merits of his views. In so doing, Kahn realized that the traditional substantive failings of the CAB went hand-in-hand with addressing its procedural flaws and legacy ways of doing business—by restricting competition and awarding benefits to a precious few, it was natural for the agency to operate along political lines as opposed to judicial ones.[13]

At the level of strategy, whether conscious or not, Kahn's overarching objective was to help foster a political environment conducive to regulatory reform. As Michael Levine (2006) describes it, the central challenge that Kahn faced was to overcome the presence of "slack" enjoyed by the legacy regulatory regime. On Levine's (2006) account, slack is the dynamic that creates political cover for regulatory beneficiaries to continue enjoying the benefits accorded to them while the public and their elected representatives are unable to appreciate the economic consequences of regulatory policy decisions.[14] In an alternative conception of slack, Professor Croley (1998, p.23) explains that "principal-agent 'slack'" is what enables "legislators [or regulators] to 'shirk'—to sacrifice their constituencies' collective interests for their own private goals and interests." On both explanations, slack represents the lack of awareness and attention to substance of the relevant regulatory policies and thus protects the beneficiaries of regulation and stands in the way of regulatory reform.

The reduction of slack is not necessarily something that a single political entrepreneur can accomplish. Rather, as Levine (2006, p.272) puts it: "Regulators can increase or minimize the transaction and information costs involved in monitoring and changing policy, and actors like media, academics, policy entrepreneurs, lobbyists, public interest groups, and politicians looking for an issue can increase or lower these costs through their own behavior." In the case of airline regulation, Kahn benefitted from and entered the

---

[12] In a famous speech to the American Bar Association, Kahn related that the traditional regulatory process involved:

> [A] lawyer on the General Counsel's staff, amply supplied with blank legal tablets and a generous selection of clichés—some, like "beyond-area benefits," "route strengthening" or "subsidy need reduction," tried and true, others the desperate product of a feverish imagination—would construct a work of fiction that would then be published as the Board's opinion.

(McCraw, 1984, p.286). This observation anticipates a later scathing critique of FCC regulation by Judge Richard Posner—"the [FCC's] opinion, like a Persian cat with its fur shaved, is alarmingly pale and thin." Schurz Comm, Inc. v. FCC, 982 F.2d 1043, 1050 (7th Cir. 1992).

[13] Kahn is quoted as stating that "[t]he dispensation of favors to a selected few is a political act, not a judicial one (McCraw, 1984, p.286)."

[14] Levine (2006, p.273 n.9) credits Kalt and Zupan for developing the concept.

surge of congressional interest in the topic that, in turn, "resulted in intense media coverage of the issue from 1976-1978 (Levine and Forrence, 1990, p.195)." Nonetheless, even if Kahn, acting alone, might have failed to spur regulatory reform, it is clear that his actions at the head of the CAB proved to be the lynchpin for regulatory reform. In particular, three distinct aspects of his leadership helped to catalyze the historic changes that occurred under his watch.

First and foremost, Kahn embraced the role of public advocate for regulatory reform. To be sure, he was fortunate to operate in a climate where other actors were interested in raising awareness about the benefits of deregulation, but his leadership—described by Levine (2006, p.277) as "intellectually powerful, attractive, and mediagenic"—was invaluable. In particular, his credibility stemmed from a widespread confidence that he was not trying to advance any self-serving political or parochial agenda. To that end, his prior academic writings—notably, his calls for airline deregulation in his Economics of Regulation—made him as bullet proof as a Washington inside player can be. And Kahn was not afraid to step into the line of fire, either, as he "testified before Congress, lobbied the White House, engaged the industry in debate, and was very accessible to the media (Levine, 2006, p.277)." Thus, a considerable part of Kahn's success as a political entrepreneur stemmed from his ability to translate his sophisticated academic understanding of regulation in terms that overcame the general bias that regulatory policy issues are too complex, esoteric, or boring to sustain public attention and concern.[15] Praising his public relations skills and ability to come up with the right one-liner, Thomas McCraw (1984, p.223) called him the "model of the modern media general."

A second component of Kahn's political entrepreneurship was his willingness to use the CAB's legal authority to move the industry toward deregulation. Notably, the agency took a number of steps to loosen its control over the industry and encourage entry,[16] thereby both facilitating experiments that made the case for further deregulation, undercut the industry's firm and united stance against deregulation, and made the ultimate move to complete deregulation less drastic for the industry players. Notably, under Kahn's leadership, the agency began, for the first time, to approve and promote the use of discount airfares, which facilitated price competition and improved dramatically the load factors on airplanes. (McGraw, 1984, p.275-77).[17] Indeed, when Congress enacted its legislation deregulating the industry, it did so recognizing that the law might actually be able to constrain the ongoing deregulatory zeal Kahn had brought to the CAB.[18] For Kahn, a willingness to use novel strategies was hardly new; at the New York Public Service Commission, he did just that, pioneering what might be called a "negotiated rulemaking," where he told the parties "I hope we can proceed as cooperative seekers of the truth, rather than as adversaries seeking an advantage over one another." (McCraw, 1984, p.251).

[15] As Levine and Forrence (1990, p.190) describe, "Most regulatory issues are highly specialized and do not generate much general attention. Complexity and lack of salience make efforts to overcome slack on regulatory issues—especially issues of economic regulation—particularly costly."
[16] See, for such an example, Oakland Serv. Case, 78 C.A.B. 593 (1978) (final order).
[17] In fairness to Kahn's predecessor, John Robson, he did oversee the approval of "supersaver" fares on an experimental basis, but it was Kahn who pushed to allow such fares more broadly and on a permanent basis.
[18] As the conference report stated, "Congress expects the deregulation of the aviation industry to move in accordance with this legislation and not in accordance with the, perhaps, differing concepts of some members of the CAB." H.R. REP. NO. 95-1779, at 56 (1978), as reprinted in 1978 U.S.C.C.A.N. 3773, 3775.

The third and final component of Kahn's strategy was to identify and bolster the interests and forces vested in deregulation. As Kahn (1990, p.331) explained, one of his goals at the CAB was to bolster the "vested interests in deregulation" on the view that the successful upstarts "having now achieved freedom from regulation, will not readily surrender it." In general, such interests are less well organized than the established interests that benefit from the status quo, meaning that they are uniquely apt to be activated by a political entrepreneur. Again, Kahn had tested this approach out in New York State, where he enlisted a number of groups (including less likely ones like the Environmental Defense Fund as well as a number of utilities) to support his initiative to reform utility rate structures (McCraw, 1984).[19] In the case of airline deregulation, it was his effort to empower and highlight upstarts like Southwest Airlines as well as to work with United Airlines, which had begun to doubt the wisdom of the legacy regulatory regime, that helped to lay the groundwork for regulatory reform.

3 Applying the Political Entrepreneurship Model to Spectrum Regulation

In many respects, airline regulation and spectrum regulation share some basic characteristics. On some accounts, both regulatory regimes emerged from an effort to protect established interests—the major airlines and the broadcasters—and limited output by restricting the use of the resource in question (Hazlett, 2001 and Keyes, 1951). In both cases, moreover, early academic criticism—famously, Ronald Coase (1959) in the spectrum context—called for regulatory reform that went unheeded (Keyes, 1949).[20] In the case of spectrum regulation, the movement toward regulatory reform came later, was halting, and remains incomplete. Indeed, Tom Hazlett has criticized the FCC's failure to escape the public choice trap of protecting incumbents by suggesting its initials can best be understood as "Forever Captured by Corporations" (Clark, 2005 and Hazlett, 2001). Part of the challenge facing spectrum reform, like the challenge faced by Kahn at the CAB, is that the FCC itself is not institutionally designed to, among other things, manage spectrum rights in a deregulated, property rights-like manner and thus invites the type of public choice behavior decried by Hazlett.[21]

In the case of spectrum reform, the biggest single force driving regulatory reform and challenging the status quo was the introduction of spectrum auctions. Before 1993, Congress called for the regulation of spectrum as a public trust whereby the Federal Communications Commission would hold comparative hearings to determine what

[19] "As McCraw (1984, p.255) reported, Kahn "[w]ork[ed] assiduously over a two year period . . . [to] shape[] a political consensus among unusual bedfellows: environmentalists, electric utilities, consumer groups, business firms, and residential customers."

[20] For an interesting study of the progress (and relative lack of progress) in the spectrum reform context, see the special issue of the *Journal of Law and Economics*, Vol. 41, No. 2, Part 2 (Oct. 1998), which emerged from a conference on "The Law and Economics of Property Rights to Radio Spectrum," sponsored by the Program on Telecommunications Policy, Institute of Governmental Affairs, University of California, Davis.

[21] The FCC's institutional limitations do not only restrict its ability to oversee property rights in spectrum. It also, as I have explained elsewhere, restricts its effectiveness in facilitating the use of unlicensed (or commons) spectrum. See Philip J. Weiser & Dale N. Hatfield, Policing the Spectrum Commons, 74 Ford. L. Rev. 663 (2005).

individual or firm merited a license to use the radio spectrum.[22] In 1993, Congress authorized the FCC to use auctions to assign spectrum licenses and, after the FCC held its first such auction (to assign new blocks of spectrum to be used for second generation (i.e., digital) cellular telephony), it generated billions of dollars in revenue for the Treasury.[23] Congress, recognizing easy money when it saw it, passed a new law requiring that all future assignments of the right to use radio spectrum be determined by an auction that would generate funds for the treasury.[24]

The radio spectrum remains, on almost any account, vastly underutilized. Thus, at the same time when entrepreneurs and users of all stripes are interested in gaining access to spectrum—and many established firms recently paid, in total, $19 billion for licenses that will become available after the digital TV transition—there is precious little spectrum available. This state of affairs reflects, among other things, that much of the wireless spectrum was allocated to specified uses well before the introduction of auctions and the FCC has historically done very little to ensure that spectrum is being used at all. To its credit, the FCC has authorized the leasing of spectrum licenses.[25] To date at least, a secondary market in spectrum has developed relatively slowly for bands in which it is authorized and, for many bands of spectrum, spectrum leasing is not allowed at all.

In 2002, the FCC, under Chairman Powell's leadership, sought to reform spectrum policy and, among other things, take aim at the agency's highly conservative and relatively undefined approach to spectrum property rights.[26] Under the agency's legacy model of regulation, licensees are protected against interference through a set of before-the-fact protective regulations that minimize the possibilities of interference. This means, among other things, that parties seeking additional flexibility must ask the agency to grant them rights to use its spectrum more flexibly—even if it highly confident (and would put its money on it, say, in a performance bond) that the different use would create a negligible, if any, amount of additional interference. During his tenure, Powell did not see any of these initiatives to completion and his successor declined to spearhead them, closing the relevant proceedings without attempting to address the underlying issue.

Powell's spectrum reform effort is notable both for its creative thinking about the need to define property rights in spectrum as well as the lack of any effort to re-think current spectrum allocation decisions. Consider, for example, that Powell did not attempt to develop a framework for enabling UHF TV broadcast licensees to allow other users (such as wireless broadband providers) to gain access to such spectrum.[27] Similarly, Powell did

[22] This system, as public choice theory would predict, rewarded politically connected firms and made entry more difficult. Among the famous tales of public choice behavior were Lyndon Johnson's ownership of radio and TV stations in Texas (Lady Bird Johnson was awarded the licenses). For a discussion of this episode, and how spectrum licenses were historically managed, see Nuechterlein & Weiser (2005).

[23] For the auction revenues collected see CAPCP, FCC Spectrum Auction Data, available at http://econ.la.psu.edu/CAPCP/FCC/index.html.

[24] Omnibus Budget Reconciliation Act of 1993, Pub. L. No. 103-66, § 6002 (1993).

[25] Promoting Efficient Use of Spectrum Through Elimination of Barriers to the Development of Secondary Markets, Second Report & Order, Order on Reconsideration, & Second Further Notice of Proposed Rulemaking, 19 FCC Rcd. 17,503, 17,505 (2004).

[26] The centrepiece of Powell's reform program was set out in an interagency task force report. See FCC, Spectrum Policy Task Force Report, ET Dkt. No. 02-135 (2002), available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-228542A1.pdf.

[27] To be fair, Powell's focus on facilitating the digital TV transition may well have limited his willingness to push on this point. For a discussion of a proposal aimed at addressing this very issue, see Philip J. Weiser,

not press—in large part because the FCC lacks jurisdiction over the issue—the question of whether the federal government is effectively using its spectrum.[28]

Future Commissions will confront the opportunity to advance the cause of spectrum reform along a number of possible dimensions. First, as noted above, there needs to be greater scrutiny of whether legacy allocations still make sense in light of the current demand for spectrum for emerging uses, such as wireless broadband. Second, the legacy approach for defining rights to use spectrum is, on all accounts, highly conservative and prevents spectrum from being used more intensively.[29] Third, there is a huge opportunity to welcome and encourage the development of new technologies, such as "smart radio" technologies, that can provide for more efficient uses of spectrum. Finally, there needs to be greater transparency of what spectrum licenses are out there and how they are being used.

One obvious question is how a future political entrepreneur at the FCC could follow Kahn's example to spearhead successful spectrum policy reform. The threshold and most fundamental challenge that a political entrepreneur would face in spurring spectrum policy reform along the lines discussed above is that the issue has proven politically intractable and incomprehensible to many legislators, let alone, the public at large. The formidable nature of this challenge has led some commentators, such as Jim Speta (2005), to doubt whether spectrum reform is readily achievable.

A political entrepreneur interested in developing the public case for spectrum policy reform confronts two notable challenges. First, many academics continue to debate different strategies (notably, whether to emphasize property rights or unlicensed spectrum), the appropriate tactics necessary to advance a particular strategy (say, whether to use courts or an administrative agency to oversee the use of spectrum property rights), and the proper role of spectrum allocations (what amount of spectrum should continue to be dedicated to over-the-air broadcast, for example).[30] Second, the entire vocabulary and discourse around spectrum policy is arcane and intimidating to outsiders. Thus, this state of affairs is a product of the culture of telecommunications regulation more broadly where, as Nicholas Lemann (2002) once explained, it is "an insider's game, less because the players are secretive than because the public and the press—encouraged by the players, who speak in jargon—can't get themselves interested."

The Untapped Promise of Wireless Spectrum, The Hamilton Project (2008), available at http://www.brookings.edu/~/media/Files/rc/papers/2008/07_wireless_weiser/07_wireless_weiser.pdf.

[28] In a later initiative, the Office of Management and Budget did just that, mandating that government users create and use a "shadow price" that the captures the value inherent in their use of spectrum with the aim of dissuading them of the fallacy that their use of spectrum is free. *See* Office of Management and Budget, "Preparation, Submission, and Execution of the Budget Section 33." Circular No. A-11 (2007), available at http://www.whitehouse.gov/omb/circulars/a11/current_year/a_11_2007.pdf. As of this writing, however, it remains to be seen how effectively this requirement will be enforced.

[29] For a critique of this system, see Philip J. Weiser & Dale Hatfield, Spectrum Policy Reform and the Next Frontier of Property Rights, 15 Geo. Mason L. Rev. 549 (2008). For suggestions of the role that the FCC should play with respect to unlicensed spectrum, see Philip J. Weiser & Dale N. Hatfield, Policing the Spectrum Commons, 74 Ford. L. Rev. 663 (2005).

[30] Jim Speta (2005 p. 185 n. 12 & 186) discusses the literature of different strategies for spectrum reform. For a sense of the debate on appropriate ways to implement spectrum property rights, compare Philip J. Weiser & Dale Hatfield, Spectrum Policy Reform and the Next Frontier of Property Rights, 15 Geo. Mason L. Rev. 549 (2008) with Thomas W. Hazlett, A Law & Economics Approach to Spectrum Property Rights: A Response to Professors Weiser & Hatfield, 15 Geo. Mason L. Rev. 975 (2008).

For those not well versed in the jargon of spectrum policy, it is difficult to appreciate how the current regulatory discourse clouds the relevant policy stakes of spectrum regulation. As J.H. Snider (2007) highlights in his article, *The Art of Spectrum Lobbying*, there is an enormous amount of information related to spectrum licenses—including the basic fact about who holds what licenses—that is not publicly available or, when theoretically available, is discussed in terms that are incomprehensive to anyone other than those very close observers of spectrum policy. The net result of this state of affairs is that the intellectual barriers to entry for understanding spectrum policy are seen as great and the benefits of trying to overcome them are seen as not worth the candle.

Assuming that a political entrepreneur can develop a compelling and popularly accessible rationale for spectrum reform, Kahn's example suggests a few additional prerequisites for the successful implementation of spectrum policy reform. First, it will be important that congressional leadership—like the kind demonstrated by Senator Kennedy with respect to airline deregulation[31]—emerges to champion the issue. Any such effort will need to pick up on the efforts of others to craft a simple narrative—lower prices or more advanced services for consumers, for example—that can be easily understood by policymakers and the public. Second, a political entrepreneur will need to look for ways to use the FCC's existing authority to pursue reform. Consider, for example, that the FCC has considerable legal latitude to reorient some of its own approaches for defining and enforcing property rights in spectrum.[32] Third, a political entrepreneur will need to develop support from those likely to benefit from future reform in this area.

Even if a future President taps an FCC Chair who pursues spectrum policy reform with the same level of zeal, political deftness, and intellectual honesty exhibited by Kahn in the case of airline deregulation, she will not succeed unless the overall political climate is receptive to reform. The public's level of interest in an otherwise arcane topic depends and openness to reform depends not merely on the development of an accessible narrative, but also a sense of crisis that renders the narrative compelling and important. With respect to airline deregulation, for example, the presence of rampant inflation provided an impetus for examining regulatory reform strategies that promised to lower the prices of services available to consumers.

The final ingredient of implementing spectrum reform also draws inspiration from Kahn's leadership at the CAB. In particular, if such an opportunity arises, a critical challenge—not unlike Kahn's challenge at the CAB—will be to reform how the FCC operates (i.e., its "spectrum culture"). At present, the agency focuses almost entirely on avoiding any potential for interference between users of spectrum, generally relying on the *ex ante* strategy of instituting protective measures. In so doing, it almost always avoids evaluating the actual uses of spectrum *ex post*. In this respect, spectrum regulation still largely reflects the command-and-control legacy of an earlier era and has yet to transition fully to a new model of regulation that would make use of adjudication proceedings that evaluate whether interference occurs in practice (and not merely in theory).

---

[31] For a discussion of the importance of the Kennedy hearings, see John W. Kingdon, A Model of Agenda Setting, With Applications, 2001 L. Rev. Mich. St. 331.

[32] For one such example, consider the FCC's reform of its rules governing broadband over powerline systems, where it instituted an after-the-fact assessment strategy in lieu of the traditional use of protective rules. *See* Amendment of Part 15 Regarding New Requirements and Measurement Guidelines for Access Broadband Over Power Line Systems, *Report & Order*, 19 FCC Rcd 21,265, 21,291-96 (2004).

4 Conclusion

In this Essay, I have sought to highlight that regulators have a lot to learn from Fred Kahn. As a political entrepreneur, he informs both the theory and practice of regulation like few others. As a scholar, he has called for evidence-based policy analysis that, on account of his public service, he has demonstrated can be put into practice. In so doing, he has challenged public choice theory's strongest claims and has highlighted that calls for regulatory reform are not merely tilting at windmills. Despite these accomplishments, scholars have yet to pay sufficient attention to different case studies of political entrepreneurship and Fred Kahn's leadership in particular. By sketching out how regulatory reform in the airline industry reflects the success of Kahn (and others) as a political entrepreneur, this Essay underscores that such studies are worth pursuing.

In closing, I must also acknowledge a quality of Fred's that enabled him to succeed and make him such a special role model—his concern with and engagement with others. Fred is a social person and his interest in a younger generation of regulators and scholars has benefitted numerous individuals (myself among them). For all of those who have gotten to know Fred, his striking combination of a willingness to learn from others, true humility, and superb intellect reminds us all not to take ourselves too seriously.

It is all too easy to despair about the state of our politics in general and regulatory reform efforts in particular. But Fred reminds us that such challenges are soluble and that we should undermine the power of ideas and determined leadership. He also leaves us all with the hopeful proposition that we can contribute as much intellectually at any age as he does at 90!

5 References

Clark, D. (2005). "Industry Experts Disagree on Best Path to Improve FCC," TechnologyDaily. March 24. http://www.nationaljournal.com/pubs/techdaily/pmedition/2005/tp050324.htm

Coase, R.H. (1959). "The Federal Communications Commission," *Journal of Law and Economics*, 2: 1-40.

Croley, S.P. (1998). "Theories of Regulation: Incorporating the Administrative Process," *Columbia Law Review*, 98:1-168.

Hardin, R. (1982). Collective Action. RFF Press: Washington.

Hazlett, T.W. (2001). "The Wireless Craze, The Unlimited Bandwidth Myth, The Spectrum Auction Faux Pas, and the Punchline to Ronald Coase's Big Joke: An Essay on Airwave Allocation Policy," *Harvard Journal of Law and Technology*, 14: 335-567.

Kahn, A.E. (1970). "The Economics of Regulation: Principles and Institutions," John Wiley & Sons, Inc., New York.

Kahn, A.E. (1971). "The Economics of Regulation Volume 2: Institutional Issues," John Wiley & Sons, Inc., New York.

Keyes, L.S. (1949). "National Policy Toward Commercial Aviation – Some Basic Problems," *Journal of Air Law and Commerce*, 16: 280-97.

Keyes, L.S. (1951). Federal Control of Entry into Air Transportation. Harvard University Press: Cambridge.

Kingdon, J. (1995). Agendas, Alternatives, and Public Policies, Harper Collins: New York.

Lemann, N. (2002). "The Chairman," *The New Yorker*. October 7.

Levine, M.E. and J.L. Forrence. (1990). "Regulatory Capture, Public Interest, and the Public Agenda Toward a Synthesis," *Journal of Law, Economy and Organization*, 6: 167-198.

Levine, M.E. (2006), "Why Weren't The Airlines Reregulated?" *Yale Journal on Regulation*, 23: 269-297.

McCraw, T. (1984). Prophets of Regulation. Belkap Press: Cambridge.

Migue, J.L. (1977). "Controls Versus Subsidies in the Economic Theory of Regulation," *Journal of Law and Economics*, 20: 213-21.

Mintrom, M. (2000). Policy Entrepreneurs and School Choice. Georgetown University Press: Washington, D.C.

Nuechterlein, J. and P.J. Weiser. (2005). Digital Crossroads. MIT Press: Cambridge.

Peltzman, S., M.E. Levine and R.G. Noll (1989). "Economic Theory of Regulation After a Decade of Deregulation," *Brookings Papers on Economic Activity: Microeconomics*, 1989: 1-59.

Pozen, D.E. (2008). "We Are All Entrepreneurs Now," *Wake Forest Law Review*, 43: 283-340.

Schuck, P.H. (1981). "The Politics of Regulation," *Yale Law Journal*, 90: 702-725.

Speta., J.B. (2005). "Making spectrum Reform Thinkable," *Journal of Telecommunication and High Technology Law*, 4: 183-215.

Snider, J.H. (2007), "The Art of Spectrum Lobbying," *New America Foundation Working Paper*, available at http://www.newamerica.net/files/WorkingPaper19_SpectrumGiveaway_Snider.pdf.

Stigler, G. J. (1971) "The Theory of Economic Regulation," *The Bell Journal of Economics and Management Science*, 2: 3-21.

Walker, J.L. (1974), "Performance Gaps, Policy Research, and Political Entrepreneurs: Toward a Theory of agenda Setting," *Policy Study Journal*, 3:112-116.

Weber, M. (1946). Politics as a Vocation. In H.H. Gerth & C. Wright Mills (eds), From Max Weber: Essays in Sociology. Oxford University Press: Oxford.

Wilson, J.Q. (1975). The Politics of Regulation. In J.W. McKie (ed), Social Responsibility and the Business Predicament. Brookings Institution Press: Washington.

Wilson, J.Q. (1980). The Politics of Regulation. Basic Books: New York.

6 Acknowledgments

Thanks to Brad Bernthal, Andrew Crain, Ray Gifford, Ellen Goodman, Jon Nuechterlein, Adam Peters, Jim Speta, Tim Tardiff, Heidi Wald, Dennis Weisman, and an anonymous referee for helpful comments and encouragement and to Dan McCormick for outstanding research support.

Public Choice has demystified and undeified the state.

The Public Choice Revolution

By Pierre Lemieux
Université du Québec en Outaouais

A half-century ago, an orthodox economist would approach the analysis of public policy with the following reasoning: Markets are efficient, or "Pareto-optimal," when perfect competition prevails. Pareto optimality means that there is no way to reallocate inputs or outputs to benefit some individual without harming another individual or, thought of another way, all gains from exchange have been realized. In many cases, such results are precluded by different types of "market failure" like macroeconomic imbalances, natural monopoly, or externalities (positive or negative). Positive externalities can be generated by "public goods," which provide benefits to everybody as long as the goods are produced and consumed by somebody. Government must intervene to correct market failures and maximize social welfare.

That was policy analysis before the public choice revolution. Today, the view is much different and begins with a simple question: How are collective decisions made? The answer, of course, is that the decisions are made by policymakers — politicians and bureaucrats — and by voters. The starting idea of public choice theory is disarmingly simple: Individuals, when acting as voters, politicians, or bureaucrats, continue to be self-interested and try to maximize their utility.

Excluding immediate precursors like Anthony Downs' 1957 book *An Economic Theory of Democracy* and Duncan Black's 1958 book *The Theory of Committees and Elections*, the foundation of the public choice school can probably be dated to the 1962 publication of James Buchanan and Gordon Tullock's *The Calculus of Consent*. Many well-known public choice economists were congregating around Buchanan and Tullock at Virginia Tech at that time: Geoffrey Brennan, Robert D. Tollison, Richard E. Wagner, Winston Bush, and others. For his seminal work in public choice, Buchanan was awarded the 1986 Nobel Prize.

In a narrow sense, public choice analysis is concerned with "state failures." Manned by self-interested actors on a "political market," the state is often incapable of correcting market failures — or, at least, of correcting them at a lower price than the cost of the original market failures themselves. In a wider sense, public choice is, as Dennis Mueller writes in his book *Public Choice III*, "the economic analysis of political institutions." In this broad sense, virtually all economists who study government intervention have now become public choice economists.

THE STATE

Why do we need the state to provide certain goods and services? Why not just have anarchy and let everyone fend for himself either individually or as a member of a private group? The subtitle of James Buchanan's seminal 1975 book *The Limits of Liberty* summarizes where the individuals presumably want to be: "Between Anarchy and Leviathan." In the mainstream public choice perspective, the state is necessary to stop the Hobbesian "war of all against all." As Mancur Olson puts it, a "sedentary bandit," the state, generates more prosperity than the "roving bandits" it puts out of business.

Once it is admitted that the state is necessary, positive public choice analyzes how it assumes its missions of allocative efficiency and redistribution. Normative public choice tries to identify institutions conducive to individuals getting from the state what they want without being exploited by it.

The contractarian approach defended by many public choice theorists is part of the normative leg of public choice theory. It distinguishes a "constitutional stage" in which, conceptually, individuals unanimously accept the rules of the political game, and a "post-constitutional stage" in which the rules of day-to-

Pierre Lemieux is an economist at the Department of Management Sciences of the Université du Québec en Outaouais, and a research fellow at the Independent Institute. He may be contacted by e-mail at pierre.lemieux@uqo.ca.

Why the Majority Wants Both A and Not-A

Assume a society composed of three voters: X, Y, and Z. Further, assume that they are deliberating over three mutually exclusive policy proposals: A, B, and C.

Table 1 shows each voter's order of preference among the three proposals. For example, Voter X's favored proposal is A; his second preference is B, and his least preferred policy is C. The other voters have different preferences. Note that we are only assuming ordinal utility, which means that only the rank of the alternatives matters.

Now, suppose that a referendum asks the voters to choose, under a majority rule, between proposals A and B. Proposal A will win because both X and Y prefer A to B and will vote consequently; Z, who prefers B, will be outvoted. If, instead, the electorate is presented with a choice between B and C, B will win because that is preferred by X and Z. Now, consider what happens if the voters are given the choice between A and C. C will win because voters Y and Z will vote for C. This means that the "social preferences" expressed by the vote are intransitive — A is preferred to B, B is preferred to C, but C is preferred to A.

This happens because the preferences of some voters (Voter Y in this case) are not "single-peaked." To see what this means, imagine a one-dimension continuum from A to B to C (like if they are placed in this order along an axis). Voter X's preferences are single-peaked because, from his most preferred alternative, A, he continuously loses utility as he moves further from A in either direction. In this case, he cannot move "left" from A because he is at the left extreme on the continuum; but the further he goes right, the less utility he gets. Similarly, Voter Z's preferences are single-peaked: his most preferred alternative is B, in the middle of the continuum, and he looses utility as he moves left to A or right to C. But Y's preferences are not single-peaked: he most prefers C and, as he moves left to B, his utility decreases, but then increases again at A. R

TABLE 1

An Illustrative Case of Voting

	First Preference	Second Preference	Third Preference
Voter X	A	B	C
Voter Y	C	A	B
Voter Z	B	C	A

day politics apply. This latter stage typically involves decisions based on majority approval, not unanimous agreement.

CYCLING Why would individuals agree to have collective choices made by majority rule? There is only one way an individual can be sure not to be exploited by a majority: have veto power over any collective choice or, in other words, require that all decisions be approved unanimously. However, unanimous agreement is practically impossible because the decision cost is prohibitively high — a larger number of individuals will have to be persuaded, and it will be in the interest of each one to lie on his preferences in order to manipulate the decision and get the highest benefits and pay the lowest taxes. On the other hand, the lower the required plurality, the higher the risk of an individual being exploited. Depending on the cost of reaching a decision and the lost benefits if the wrong decision is made, the proportion required may be a simple majority (50 percent plus one vote) or some higher, qualified majority.

Here, we meet the first problem uncovered by public choice analysis: majorities — especially simple majorities — are arbitrary. It is far from clear who the majority is and what it wants. Majorities can be inconsistent or, what amounts to the same, can cycle indecisively among alternatives. (See "Why the Majority Wants Both A and Not-A," above." The majority can vote for Proposal A over Proposal B, and B over Proposal C, but then vote for C if asked to choose between A and C. The majority seems intransitive, irrational. This explains why the electorate often appears so inconsistent — for example, when it votes for minimum wages that create unemployment, and then for government programs meant to create jobs.

Cycling can occur when the preferences of some voters are not "single-peaked" — that is, given a broad spectrum of options, the force of the voter's preference does not consistently decline as he considers options further and further from his top choice. There is nothing in individual rationality that implies single-peaked preferences. As Mueller notes in *Public Choice III*, "During the Vietnam War, it was often said that some people favored *either* an immediate pullout or a massive expansion of effort to achieve total victory." Another example: Some people think that either selling tobacco should be forbidden or else smokers should be left alone; middle-of-the-road options are less preferred (by some).

Another way to see the cycling-inconsistency phenomenon is to understand how redistributive coalitions are unstable. If the middle class votes with the poor, the two classes can expropriate the rich. But then it will be in the interest of the rich to bribe the middle class into a new coalition that advocates taking all possible money from the poor and offering a bit more to the middle class. The poor, realizing that they would be better off if they were less exploited, would then offer a new deal to the middle class to exploit the rich and share the loot differently. And so on to a new winning coalition. What, then, does the majority want?

Of course, "the middle class" does not vote like a single person and, in a simple majority system, any coalition comprising a simple majority of voters will do. Logrolling is the way coalitions will often be formed in practice. "Logrolling" describes an informal exchange of votes: Politician X supports a measure he does not much like but Politician Y finds very important in exchange for Y's support for a measure that X wants dearly. If such trade in votes is possible, somebody else can outbid one of the traders and get support for his own pet project. That new coalition can also be overturned. Mueller reports some empirical evidence that such vote trading happens in Congress — for example, lawmakers representing peanut farmers voted for initiatives favored by the sugar-farming industry in return for votes from sugar farmers' congressmen on peanut industry measures.

MEDIAN VOTER THEOREM When no cycling occurs, the median-voter theorem kicks in. In a simple one-dimensional case, the theorem states that, if all voters have single-peaked preferences, the winning alternative in a simple majority vote will be the ideal (or most favored) alternative that is at the median point of the preference distribution. This can be seen in Figure 1, which represents the preferences of voters X_1 to X_5. The horizontal axis represents the different alternatives of a one-dimensional issue — say, a smaller or larger tax rate, with the rate ranging from 0 percent to 100 percent. The strength of each voter's preference for a specific tax rate is shown by the respective inverted-U curve and measured on the vertical axis. (This is measured in an "ordinal" way, i.e., voters prefer alternatives "more" or "less" — consequently the height of each voter's utility curve does not matter.)

Consider Voter X_3. His ideal tax level is t_3, which corresponds to the top of his utility curve. The more the actual tax rate diverges from t_3, the less he likes the alternative, which is just another way to say that his preferences are single-peaked. The same is true for the other four voters. By definition, X_3's ideal alternative, t_3, is the median one among the five voters' ideal alternatives.

The median-voter theorem is now easy to understand. In pair-wise voting, there is no alternative that can win against t_3 (assuming all voters vote). Suppose, for instance, that the electorate is asked to choose between t_2 and t_3. Since t_3 is the median, the majority of voters (i.e., X_3 and the two voters to his right) prefer it to any lower rate. Remember that preferences in this example are single-peaked and that a voter prefers an alternative less the further it is from his ideal one. Note that this is true no matter how the other voters' ideal

MORGAN BALLARD

rates are distributed across the continuum — the median rate will always win.

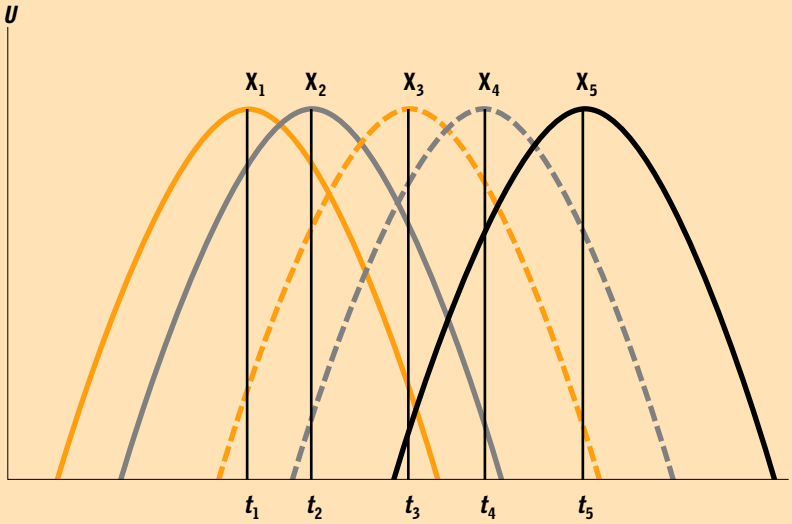
This explains why, especially in two-party systems, political platforms are so similar. If, in the upcoming U.S. presidential election, President Bush promises t_2 and Sen. Kerry promises t_5, Voter X_3 will vote for Bush because Voter X_3 gets more utility from t_2 than from t_5. Hence, t_2 will win in that pairing. If Kerry is self-interested and wants to get elected, he will move to t_4, thereby countering Bush. Bush, if he is self-interested and wants to get elected, will then move even closer to t_3, and so will Kerry. Both politicians will move toward the median point in pursuit of an electoral win.

Except for the median voter, all voters are unhappy with the results. This feature of voting (whether in elections or referenda) is inseparable from collective choices. If a collective choice decided the type of car that everyone would drive, the outcome might be the production of a Ford Taurus for everybody. The median voter would be happy (assuming the Taurus is the preferred car of the median voter), but everybody else would rather be driving another car.

FIGURE 1

Why the Centrist Wins

An illustration of the Median-Voter Theorem.



CYCLES OR TYRANNY? Cycles grow more likely as the number of possible alternatives increases and the individual preferences become more heterogeneous. Qualified majorities can reduce the probability of cycling. An "agenda setter" who decides which alternatives are put on the ballot can also reduce cycling, but he likely will set the agenda to produce an outcome that follows his own preferences. If X is the agenda setter, he will first pair Proposal B with Proposal C and send that choice to the voters. He will then pair the winner with Proposal A, which Voter X prefers. The result of that agenda is that A will win. Likewise, if the agenda setter is Y, he will make sure that the first vote pits Proposal A against Proposal B, and will then run the winner against C so that C wins.

If there is no cycling, another danger looms: a stable majority that would exploit and oppress an identifiable minority. We seem to be facing an uncomfortable choice: either inconsistent majorities with unstable coalitions, or agenda-setter dictatorship, or else a Tocquevillian "tyranny of the majority."

DIFFERENT VOTING RULES Besides the simple majority rule analyzed above, there is a large number of different voting rules meant to find out who exactly is the majority and what it wants, especially when more than two candidates are running. The plurality rule just picks the candidate who gets the highest percentage of the vote. The majority rule with runoff election pairs the two candidates who obtained the highest percentage of the vote in a first round. The Condorcet criterion calls for the winner to be the one who can defeat all the other candidates pairwise. The Hare system and the Coombs system eliminate the least-preferred candidates over many rounds. In the approval voting system, each voter votes for many candidates and the most popular, according to this criterion, wins. The Borda count asks voters to score the candidates, and the winner is the candidate with the highest aggregated score.

All of those systems have different advantages and drawbacks, and often give different results. "For a given set of voters with unchanged preferences," Gordon Tullock concludes, "any outcome can be obtained by at least one voting method." This is not without consequences in the real world: Richard A. Joslyn's research suggests that, in the 1972 Democratic primaries, Edmund Muskie would have won against George McGovern under voting procedures other than the plurality rule.

When the issue to be decided relates to the production and financing of public goods, each individual is incited to understate what he wants, hoping to get the benefits while paying lower taxes. "Preference-revealing" systems have been devised theoretically with the aim of bringing voters to reveal truly what each alternative is worth to them. Mueller is optimistic that this line of research will eventually lead to better collective choice mechanisms. But he also shows how voting with one's feet — exit as opposed to voice — is often a better solution: In a decentralized political system, individuals get the combination of public goods and taxes they want by moving to the jurisdictions that offer their preferred packages. They reveal their true preferences by choosing where to live.

WHY VOTE? In the 2000 presidential election, the voter turnout (proportion of registered voters who did cast a ballot) was 67.5 percent. For public choice theory, the problem is not that one-third of the voters did not vote, but that two-thirds did. Why did they bother? Granted, for the typical individual voter, the cost of voting is low and mainly involves the time spent going to the polls. But the expected benefit (the value of the benefits

Does Your Vote Really Count?

Although the question of the decisiveness of a vote rapidly gets technical, the general ideas can be grasped easily.

Suppose that you are a member of a committee of three persons, X, Y, and Z, including yourself (Y). A simple majority vote is to be taken on Proposal A vs. Proposal B. Assume that there is a 50–50 chance that any of the other two members will vote for A, and likewise for B. There is no abstention or third alternative. What is the probability that your vote will count, in the sense of being decisive? This amounts to asking the probability that there would be a tie if you did not vote.

Because Voters X and Z can each vote for either A or B, there are four possible outcomes of the vote without your voice: A-A (i.e., Voters X and Z both vote for Proposal A), B-B, A-B, and B-A. Only in the last two (A-B and B-A) is there a tie without you; in the other cases, it makes no difference whether you vote for A, for B, or do not vote at all. Therefore, there are two chances in four that your vote will count.

Those who have done some elementary combinatorial analysis and probability theory will recognize that if p is the probability that one of the other voters chooses for A, and $(1\text{-}p)$ the probability that he votes for B instead, then the probability P that there will be a tie is given by the formula

$$P = {}_nC_{n/2} \times p^{n/2} \times (1-p)^{n/2}$$

where ${}_nC_{n/2}$ is the number of combinations of all n voters taken $n/2$ at a time. In fact, with a large number of voters (1,000 or more in the table), an approximation to the above formula has to be used to calculate P, even with computers.

By counting the possible results, calculating their equal probabilities, and adding the probabilities for the tied results, or alternatively by using the above formula, it can be calculated that on a five-person committee (including you), the probability that your vote will count is ⅜ or 0.375.

The larger the number of voters, the lower the probability that your vote will count. With 1,000 voters, the probability is only about 0.02 (or 1:50). With 100 million (the approximate number of voters in a presidential election), it is reduced to 0.00006 (about 1:17,000).

The probability that your vote will be decisive diminishes dramatically as the probability that any voter chooses one alternative or the other diverges from 50–50. This is easy to see in our example of a three-member committee. Assume now that each of the other two members votes for A with probability 0.8 (or, alternatively, votes for B only once in five times). Now, the two results that make your voting worthwhile (AB and BA) only occur 4/25 of the time each, so that the occurrence of one or the other event has a probability of 8/25, or 0.32.

Consider again the upcoming presidential election. Assume that about 100 million other electors will vote, and that the probability that one of those voters taken randomly will vote for Kerry is 0.49 and the probability he will vote for Bush is 0.51. Disregard the complication of the electoral college. The probability that there will be a tie, and that your vote will make a difference, can be calculated: it is eight chances in 10^{8691}. The number 10^{8691} (10 followed by 8,690 zeros) is a more-than-astronomical number. It is much, much larger than the total number of elementary particles in our universe (10^{100}) and the age of the universe in seconds (3×10^{17}). Consequently, the probability that your vote will count is extremely close to zero.

There have been some recent challenges (mentioned by Mueller in *Public Choice III*) to this methodology, and some higher probabilities have been obtained. But, for all practical purposes, one vote — your vote — still has no significant probability of changing the election outcome. R

TABLE 2

Probability that Your Vote Will Count

Number of electors (n) without you	Probability that an elector votes for one of two alternatives (p)	
	$p = 0.5$	$p = 0.49$
2	0.5	0.4998
4	0.375	0.3747
1,000	0.0189	0.0155
10,000,000	0.0002	$5 / 10^{873}$
100,000,000	0.00006	$8 / 10^{8691}$

Note: Probabilities for n = 2 and n = 4 are exact; for other values of n, they are approximated by Owen and Grofman's formula as modified by Mueller (cf. Public Choice III, p. 304–305).

promised by the voter's preferred candidate multiplied by the probability that the voter's action will get the candidate elected) is infinitesimally small. Consequently, the rational voter should abstain. Why so many voters go to the polls is called "the voting paradox."

In his new book, Mueller reviews many solutions that have been proposed to the voting paradox. The common idea is that self-interest has to be conceived in a less narrow fashion: the voter must consider other benefits than the expected gains from his preferred candidate's policies. Those other benefits might be the pleasure of expressing an opinion or of being part of the crowd (like when one cheers or boos at a hockey game), the satisfaction of doing one's moral duty, or the desire to be seen as a "good citizen." Indeed, there is some evidence that people often vote against their own narrow interests. In brief, the benefits individuals get from voting are related to their

expressing their "public preferences."

The voter weighs those benefits, which are independent of his being on the winning or losing side, against the cost of voting. The higher the cost of voting (other things being equal), the lower the turnout. Mueller reports research showing that, before poll taxes were made illegal in Southern states, "a six-dollar poll tax in 1960 reduced the probability of an individual voting by 42 percent." Voters are rational — they vote for other reasons than expected benefits from their preferred candidates or measures, but they weight those reasons against the cost of voting.

This conclusion has massive consequences. One is that, because a voter's expression has little or no impact on the election results, he has no incentive to be well-informed on political issues. Because voters thus remain "rationally ignorant," politicians can, to some extent, cheat on their commitments and indulge in their own personal ideological preferences.

DISTORTING INTERMEDIARIES

In our democracies, voters do not decide most issues directly. In some instances, they vote for representatives who reach decisions in parliamentary assemblies or committees. In other instances, they elect representatives who hire bureaucrats to make decisions. The complexity of the system and the incentives of its actors do not necessarily make collective choices more representative of the citizens' preferences.

POLITICIANS Public choice theory assumes that politicians want to win elections — otherwise they will not be politicians for long. To achieve their goal, the politicians propose measures that they think the majority prefers, and they join political parties. "Parties formulate policies in order to win elections," wrote Anthony Downs in *An Economic Theory of Democracy,* "rather than win elections in order to formulate policies."

The rule used for U.S. congressional elections and parliamentary elections in most British-tradition countries is plurality-based. This encourages the dominance of two political parties because third parties can get few candidates elected even if they can get a good proportion of the vote across all districts. In two-party political systems, as long as issues are one-dimensional and preferences are single-peaked, political parties will try to get as close as possible to the median voter. In multidimensional cases or with multiple-peaked preferences, cycling can occur.

In political systems with proportional representation, more than two political parties will usually thrive and will form coalitions to govern. Different systems of proportional representation exist, but the general idea is that some adjustment procedure brings the proportion of representatives closer to the proportion of the popular vote received by different parties in the whole country. Proportional representation gives better representation to voters with minority preferences or ideologies, but it also leads to more unstable governing coalitions.

BUREAUCRATS When public choice analysis is applied to the bureaucracy, it uncovers other reasons to doubt that the state can efficiently reconcile the individuals' preferences and aggregate their demands for public policies. Once again, public choice assumes that the bureaucrat is an ordinary individual who, like everybody else, tries to maximize his utility.

What does this mean in practice? What does the bureaucrat do to maximize his utility, given the constraints placed upon him? Most attempts to answer that question are adaptations of the systematic model William Niskanen developed in his 1971 book *Bureaucracy and Representative Government.* Bureaucrats are assumed to maximize the size of their bureaus' budgets because they can thereby increase their real remuneration in terms of perks (larger offices, better expense accounts, etc.), lower risk of missing their objectives, recognition, etc. Thus, the bureaucrats will produce more than the politicians (and, presumably, the citizens) want, or at a higher cost.

Bureaucrats can do this because their political sponsors (the politicians or other political masters who determine the agencies' budgets) do not know the real cost of producing what they order from the bureaucracy. Of course, the sponsors will try to control bureaucratic activities, but the monopolistic (single-provider) advantage of the bureaus will defeat at least some of those efforts. This theory is supported by a large number of studies showing that production costs in public bureaucracies are higher than in comparable private enterprises.

Another way the bureaucrat exerts power is by being an agenda setter. As we have seen, the agenda setter can often lead the system toward the results he prefers by deciding which alternatives, in what order, will be voted upon by the politicians or the citizens.

INTEREST GROUPS In order to influence collective choices, citizens have to engage in collective actions: participate in demonstrations, organize lobbying activities, contribute to political parties, etc. The result of a group's collective action (say, tariff protection) is often a public good for the members of the group — each will benefit from it whether or not he has contributed. Moreover, the action of a single individual may not count much in the final success of the collective action. Thus, in participating in a collective action, every individual incurs costs for virtually no benefit and will therefore be tempted to free ride on the efforts of others. That is especially true in large groups where one individual's actions have less impact and shirkers can more easily avoid sanctions like boycott. Thus, small groups with concentrated interests like farmers or steel producers will be better organized and wage more efficient collective action than large groups with diffuse interests like taxpayers or steel consumers. And well-organized groups will exploit less organized ones.

That was a major conclusion of Mancur Olson's highly influential book *The Logic of Collective Action.* For example, Swiss farmers, who are a small part of their country's population, get effective subsidies equivalent to 86 percent of their incomes while farmers in Ghana, who are a large part of their country's population, get an effective negative subsidy amounting to 27 percent of their incomes as they subsidize the small urban class.

Interest groups will engage in what public choice theorists call "rent seeking," i.e., the search for redistributive benefits at the expense of others. The larger the state and the more

benefits it can confer, the more rent-seeking will occur. "The entire federal budget," writes Mueller, "can be viewed as a gigantic rent up for grabs for those who can exert the most political muscle." Rent seeking does not produce a pure transfer; when individuals or groups compete for some advantage from the state (e.g., a subsidy, a monopoly), they will all use real resources (e.g., ink, paper, travel, meals, time) in trying to grab it. As a result, part of the expected rent will be dissipated, creating a net social loss.

GROWTH AND JUSTIFICATIONS OF THE STATE

From about 1870 until around 1913, total government (federal plus state) expenditures in the United States was around 7 percent of GDP. It grew to 12 percent by 1920, and to 20 percent by 1937. By 1980, it had reached 31 percent and has slightly increased since then. The same evolution happened in the public sectors of other countries, though many of them have seen even more growth post-1980, and some now are as high as, or higher than, 50 percent of GDP. Why did the state grow so much?

TWO MODELS Public choice theory oscillates between two models of the state: the citizen-demand model and the Leviathan model. In the citizen-demand model, Mueller explains, "policies are reflections of the preferences of individual voters." In the Leviathan model, he says, "it is the preferences of the state, or of the individuals in the government, that are decisive." Depending on which model is used, analysis of why the state has grown yields different results.

The citizen-demand model claims that the citizens have demanded more public goods, more control of negative externalities (like pollution), and more redistribution of income. One interesting hypothesis, supported by some research, is that the expansion of the voting franchise to women and the poor has fueled the growth of the state. But is this really what the citizens want? And which part of the citizenry?

The Leviathan model stresses factors on the supply side: the state grows because its rulers or beneficiaries want more loot. Bureaucrats comprise a large part of the electorate — 15 percent on average in OECD countries, and 20 percent or more in some countries — and research shows that their voter turnout is higher than other citizens'. Another explanation is that politicians can easily fool rationally ignorant voters.

Mueller, in *Public Choice III,* cites evidence that direct democracy and federalism provide an effective check on the growth of Leviathan. This suggests that the citizen-demand model is more applicable in such cases.

SOCIAL CHOICE THEORIES What are the standards for judging state action? What should the state do, and how? Those questions are the subject matter of normative theories of public choice, which include "constitutional economics" as well as "social choice" theories outside the narrow public choice school.

In social choice theories, we meet the sort of "social welfare functions" that preceded the public choice revolution. A social welfare function is a ranking, in terms of social welfare, of all Pareto-efficient configurations of prices, wages, and income distributions. To choose between them, "society" — which in practice means the state — must weight the utility of different individuals. Before the development of the public choice school of thought, it was already known that social welfare functions require that utility be measurable in a cardinal sense and be comparable between individuals. In other words, somebody must calculate that, say, a 10 percent increase in X's utility is worth "to society" more than a 15 percent drop in Y's utility. There is no scientific basis for such moral judgments or for the social welfare functions based on them.

> The entire federal budget can be viewed as a gigantic rent up for grabs for those who can exert the most political muscle.

In his 1951 book *Social Choice and Individual Values,* Kenneth Arrow attacked the problem with a different methodology. He asked if a social welfare function could be built from a few simple axioms, avoiding any interpersonal comparison of utility. In trying to answer the question, he demonstrated that no social welfare function exists that will, at the same time, respect Pareto optimality, be nondictatorial, and be transitive (i.e., not lead to cycles). In a sense, Arrow generalized the conclusions on cycling and showed that social preferences must be either inconsistent or dictatorial.

TAMING LEVIATHAN

There is an ambivalence in the corpus of theories and empirical evidence called "public choice" that Mueller encapsulates well in *Public Choice III:*

> Some scholars like Brennan, Buchanan, Niskanen, and Usher look at the state and see a grasping beast set upon exploiting its power over citizens to the maximum degree. Others, like Breton and Wittman, when they gaze upon the state, see an institutional equivalent to the market in which democratic competition produces

efficiency levels comparable to those achieved by market competition.

Public choice can be seen as demonstrating either the usefulness of politics or the existence of state failures; as arguing either that the state promotes allocative efficiency or that it is a redistributive machine; as proposing either a demand model of the state where political competitors respond to citizens' demands or a supply model where Leviathan rules. Expressed differently, there is a permanent tension between the libertarian and the interventionist strands in public choice theory.

One thing is sure: public choice has destroyed the naïve view that, in order to justify state intervention, it suffices to show that there exist market failures that an ideal state could correct. After the public choice revolution, political analysts cannot be satisfied with comparing real markets with an ideal state; they must analyze the state as it is before dreaming about what it should be. Public choice has demystified and un-deified the state.

GOING FURTHER Dennis Mueller's *Public Choice III* provides a masterly overview of the public choice literature over the last half century. It will remain as the *Summa Theologica* of public choice economics at the beginning of this century.

In my view, we can go further than Mueller, who perhaps tends to be overly optimistic about politics. The more extended is the state's domain, the more likely its majorities will be either inconsistent or oppressive. The paradigmatic case is perhaps the cycling phenomenon: the more the state intervenes, the more the issues will become multidimensional, the less homogeneous will the voters' preferences be, and the more inconsistent or dictatorial Leviathan must become.

We must question the orthodox approach of weighting costs and benefits in search of optimality. Optimality can be defined in such a broad way that everything is optimal, in the sense that it cannot be otherwise given the present state of the world. Is this not what some economists do when they argue that political competition between interest groups produces optimal results because individuals will get organized when the cost of oppression gets too high? Do we not have a tautological notion of optimality?

Thoughtful defenders of the interventionist strand do have an external criterion of optimality that lies in some sort of social welfare function. What is optimal is what maximizes some function of individual utilities, perhaps represented by monetary costs and benefits. On this basis, it can theoretically be shown that some (competitive) institutions lead to Pareto optimality and the state can use cost-benefit analysis to get us there. But, as public choice analysis has shown, this approach presupposes either interpersonal comparisons of utility or "social preferences" that are inconsistent or dictatorial.

We should take seriously the challenge raised by Anthony de Jasay, himself an (unorthodox) public choice theorist. In his book *The State,* he reminds economists that interpersonal comparisons of utility "are merely a roundabout route all the way back to irreducible arbitrariness, to be exercised by authority." "At the end of the day," he continues, "it is the intuition of the person making the comparison which decides, or there is no comparison.… In an analogous manner, the two statements 'the state found that increasing group *P*'s utility and decreasing that of group *R* would result in a net increase of utility' and 'the state chose to favor group *P* over group *R*' *are descriptions of the same reality.*" Once that is realized, public choice becomes essentially an indictment of the state.

Finally, taking public choice theory seriously implies a serious questioning of the "we" or "they" as political collectives. Statements such as "We, as a society, think or do such and such" and "The French (or the Americans) believe or do this or that" are either rhetorical and logically meaningless or else dictatorial. There is no nondictatorial way to aggregate different individual preferences and fuse them into one set of super-preferences, except if the individuals have identical preferences or if they are unanimous. Examples of similar preferences are found in small and close groups like a couple, a family, or a few friends —although even in those cases, one individual is often the "dictator" or the leader. Examples of unanimous choice include shareholders buying into a company and members joining an association. Unanimity is the only way out of the dilemma between meaningless and dictatorial collectives.

Thus, except in an abstract constitutional perspective (agreement on very basic rules), the political "we" implies that some individuals impose their preferences on others. In this sense, the public choice revolution rings the death knell of the political "we." R

READINGS

- "Dictatorship, Democracy, and Development," by Mancur Olson. *American Political Science Review,* Vol. 87, No. 33 (1993).
- *Government Failure: A Primer in Public Choice,* edited by Gordon Tullock, Arthur Seldon, and Gordon L. Brady. Washington, D.C.: Cato Institute, 2002.
- *The Limits of Liberty: Between Anarchy and Leviathan,* by James M. Buchanan. Chicago, Ill.: University of Chicago Press, 1977.
- *The Logic of Collective Action,* by Mancur Olson. Cambridge, UK: Cambridge University Press, 1965.
- *The Power to Tax: Analytical Foundations of a Fiscal Constitution,* by James M. Buchanan and Geoffrey Brennan. Cambridge, UK: Cambridge University Press, 1980.
- *Public Choice III,* by Dennis C. Mueller. Cambridge, UK: Cambridge University Press, 2003.
- *The State,* by Anthony de Jasay. Oxford, UK: Blackwell Publishing, 1985.

Army Corps of Engineers Blows Mississippi River Levee

Written by Dave Bohon

Wednesday, 04 May 2011 14:45

The Army Corps of Engineers blew open two miles of levee along the Mississippi River in Missouri May 2, an effort that may have saved the community of Cairo, Illinois, but which flooded 200 square miles of fertile farmland, which Missouri Governor Jay Nixon called “literally the most productive part of our continent.”
The *New York Times* reported May 3 that the massive flood of river released by the Army Corps of Engineers into the 130,000 acre floodway “will soon re-enter the Mississippi near New Madrid [Missouri], through two 5,500-foot stretches blasted out over two days at the lower end of the basin, and the crest will continue to roll on, with the river expected to match or beat its previous record heights at many points along the way.”

Major General Michael Walsh of the Army Corps of Engineers predicted that his team was just at the beginning of its effort to control the river, which had reached record levels over the past few days. The *Times* reported that the Army Corps’ “decision to inundate the 130,000 acres within the spillway’s basin almost certainly saved the town of Cairo, Illinois. The river had reached a record 61.7 feet at Cairo before the explosion and was predicted to rise more than a foot further.” The *Times* quoted Colonel Vernie Reichling Jr. of the Army Corps’ Memphis District as saying that they would most likely have to “fight this river all the way down to the Gulf of Mexico. I don’t see this letting up.”

According to CNN, the first blast on May 2 at the levee near Birds Point, Missouri was followed up by a second the next day at New Madrid, Missouri, with a third planned for May 4 near Hickman, Kentucky. “The second and third blasts, downstream of Birds Point, will allow floodwater to return to the Mississippi River,” CNN reported.

While the main defenses against the Mississippi River were expected to hold, additional flooding was “likely as water backs up into the rivers and tributaries that feed into the Mississippi, and tests ‘non-federal’ levees that line those waterways,” reported the *Times*. Jeff Graschel of the Lower Mississippi River Forecast Center, a subsidiary of the National Oceanic and Atmospheric Administration, told the paper that while federal levees are strong enough to handle what the river will bring, “anything that’s non-federal is a different story.”

Although residents of Cairo, Illinois, and other towns saved by the Army’s actions were grateful, those in the flooded areas, especially those who farm the hundreds of thousands of acres of fertile land, had a different take. Farmer Bryan Feezor told CNN that the sight of his submerged fields made him sick. “Farming is all I ever have done ... and it’s under water, he said. “I really don’t know [what I’m going to do].”

“It’s a life-changing event,” one local farmer, Travis Williams, told CBN News. “My heart goes out to all the farmers who lost their land and homes.”

Even with the Army Corps’ actions, a plan Walsh said had been designed decades ago for such a contingency, and which had “operated as designed," communities and residents along the Mississippi remained in danger. The residents of Cairo remained under a mandatory evacuation, and another six other communities had been issued voluntary evacuation notices, Patti Thompson, a spokeswoman for the Illinois Emergency Management Agency, told CNN. “We’re definitely not out of the woods yet,” she said. “The levees are all very saturated right now and they’re going to continue to have a lot of pressure on them for several days.”

Even with the series of breaches, the National Weather Service continued to predict near-record flooding in southern Illinois, southwest Indiana, western Kentucky and Tennessee, southeastern Missouri, northeastern Arkansas, and even into Mississippi and Louisiana.

Execution of the plan did not come without major conflict, with the state of Missouri taking the Army Corps of Engineers to federal court over the plan. “The state argued the floodwater would deposit silt on the estimated 130,000 acres, and years, along with millions of dollars, would be required to fix the damage,” noted the CNN report. By Sunday, May 1, the case had made its way to the U.S. Supreme Court, which declined to intervene, clearing the way for Walsh to initiate what he called a “heart-wrenching” decision. “I’ve been involved with flooding for 10 years and it takes a long time to recover from something like this,” he said.

Reflecting afterward on the decision to blow the levees, and on the prospect that more would have to be destroyed before it was over, Walsh said that choosing between saving Cairo and other communities or saving the farms in the flood plain was not “easy or hard. It’s simply grave, because the decision leads to loss of property and livelihood,

either in a floodway or in an area that was not designed to flood."

It seemed clear that Walsh sympathized with those who would bear the brunt of his decision, telling one group of Missouri property owners, "I recognize all of your lives will be impacted, but these levees have never been under this pressure before."

According to the Associated Press, even those opposed to the Army Corps' actions "appreciate how Walsh — who is responsible for managing the entire length of the Mississippi River valley, from Canada to the Gulf of Mexico—has handled the situation." Even Governor Nixon, whose administration took the lead in trying to stop the levees from being destroyed, conceded that Walsh "has a very difficult decision to make relatively quickly. He understands the magnitude of the decision on his plate."

Nixon said his focus would now turn to having the Missouri levees rebuilt and to restoring the lost farmland to productivity. But officials warned that it could be early fall before the water drains from the land, and that sediment left by the river could do lasting damage to the once fertile acreage.

THE WALL STREET JOURNAL.
WSJ.com

U.S. NEWS | APRIL 18, 2011

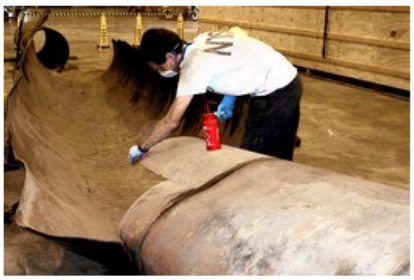
More Utilities Can't Prove Pipe Tests

By DANIEL GILBERT And REBECCA SMITH

Two more big California utilities told state regulators they couldn't prove they have done key safety tests on hundreds of miles of pipelines carrying explosive natural gas.

The disclosure late Friday by gas distribution utilities owned by San Diego-based Sempra Energy came on the heels of a similar admission last month from San Francisco-based PG&E Corp. A PG&E pipeline exploded in September, incinerating a neighborhood in San Bruno, Calif., and killing eight people.

Investigators suspect weak welds in the pipe may have played a role in the explosion.



Associated Press

A technician cleans a ruptured pipeline at a NTSB training center.

Following the San Bruno accident and urgent recommendations by the National Transportation Safety Board, California regulators in January ordered the state's gas utilities to review their safety records and report back on any gaps. The required records included pressure-test data on lines that run through populous areas; these tests are designed to flush out any leaks or flaws in a pipeline by running water or an inert gas through the line at high pressure.

In its joint report, Sempra's Southern California Gas Co. and San Diego Gas & Electric Co. said they were continuing to look, but haven't found complete pressure test records for about 450 miles—or 28%—of their pipelines that run through neighborhoods, or near parks and offices.

"We have not yet located records sufficient to document that the pipelines have been strength-tested per the NTSB recommendations," Sempra said in its filing to the California Public Utilities Commission.

Rick Morrow, the company's vice president of gas engineering, said in an interview on Sunday there were other ways to establish safe operations. He said he was confident Sempra's system was safe because 71% of gas transmission lines in populated areas, including some segments for which the company didn't have pressure-test reports, have been inspected with robotic devices called smart pigs that travel the interior of the pipes.

"I don't want people to conclude we aren't operating our lines safely because our operating history tells a lot about their safety," Mr. Morrow said.

Sempra said it would continue to search for records and in the meantime would lower the pressure at which it moved gas in some lines and increase the frequency of inspections and leak surveys.

Richard Kuprewicz, a pipeline safety consultant based in Redmond, Wash., said he was surprised that the utilities have such gaps in their records, which he said exposed a deficiency in oversight.

"The Public Utilities Commission is not regulating gas transmission pipeline safety in the state of California," Mr. Kuprewicz said in a telephone interview. "The public has a right to be concerned."

A month ago, PG&E said it had been able to find complete records for only two-thirds of its high-pressure lines that run through populated areas.

After the San Bruno explosion, federal accident investigators learned that PG&E's records wrongly identified the pipe that ruptured as seamless, an error that meant the utility never conducted the sort of inspections that might have detected a problem with seam welds in that pipe, which was more than 50 years old. State regulators want to know where similar pipe may have been used but are hitting a wall due to gaps in records.

In a potential settlement with regulatory staff, PG&E agreed to pay a $6 million fine, with $3 million suspended if it met future targets in its continuing records search. The state commission hasn't yet approved the fine and some public officials, including the state attorney general, said it was too small.

Sempra says it is in compliance with all state rules and faces no penalties. A spokeswoman for the California Public Utilities Commission said Sunday the company's filing was "under review."

THE WALL STREET JOURNAL.

Apple's Mobile Rules To Get FTC Scrutiny

By Thomas Catan

305 words
12 June 2010
The Wall Street Journal
J
B1
English

WASHINGTON -- The U.S. Federal Trade Commission will investigate whether Apple Inc.'s business practices harm competition in the market for software used on mobile devices, people familiar with the situation said.

For weeks, the FTC has been engaged in negotiations with the Department of Justice over which agency would review allegations by companies that say they're being shut-out of one of the most important emerging computing platforms.

Adobe Systems Inc. has been engaged in a public feud with Apple over its decision to ban Adobe's Flash video technology from Apple devices. This week, Google Inc. complained Apple's new rules on developers could bar Google and other rivals from selling ads inside iPhone and iPad applications, such as games.

Apple has also banned software developers from using other companies' tools to develop software for its devices.

Both Apple and the FTC declined to comment. The decision was reported earlier by Bloomberg News.

This may not be the only antitrust investigation Apple faces. Justice Department lawyers recently contacted companies about Apple's practices in the music business. It isn't clear if it would continue that inquiry.

The Justice Department is already investigating whether Apple and a range of other tech companies improperly agreed not to poach each other's employees.

The FTC will have a wealth of information to mine for its probe. It recently completed a six-month investigation of Google's $750 million acquisition of AdMob Inc., giving its lawyers knowledge of the mobile-ad market that Apple has also entered.

"The Commission has reason to believe that Apple quickly will become a strong mobile advertising network competitor," the FTC said last month.

FEATURED ARTICLE | FEBRUARY 25, 2013

The Institutions-Intensive Economy

Arnold Kling*

Today, a modern market economy with its ever-finer division of labor depends on a constantly expanding network of trade. It requires an intricate web of social institutions to coordinate the working of markets and firms across various boundaries. At a time when the modern economy is becoming increasingly institutions-intensive, the reduction of economics to price theory is troubling enough. It is suicidal for the field to slide into a hard science of choice, ignoring the influences of society, history, culture, and politics on the working of the economy.[1]

Ronald Coase, the world's oldest active economist, here aptly describes the economy in the age of the Internet, giving us the expression "institutions-intensive." His famous contributions to the theory of how bargaining might address externalities and how the boundaries of the firm depend on relative institutional effectiveness turn out to be preludes to the sort of economic analysis that is needed today.

"The plunge in manufacturing production work is indicative of the declining share of economic activity that is measurable."

Over the course of Coase's lifetime, the locus of economic activity has been shifting, from the farm to the factory floor to the office and even to "the cloud." With each step, the concept of property has become more difficult to define, the economic entities have become more difficult to locate in time and place, the proportion of wealth that is intangible has risen, and earnings have become increasingly contingent on social constructs rather than on individual attributes.

According to Bureau of Labor Statistics data, in every year between 1939 and 1956, the proportion of total employees in the nonfarm sector classified as "manufacturing production and nonsupervisory workers" was over twenty-five percent. This figure has been falling steadily, until in the last four years it has averaged just six percent.

The plunge in manufacturing production work is indicative of the declining share of economic activity that is measurable. One hundred years ago, the output of an individual farm worker was quantifiable. In principle, farm workers could be paid on the basis of specific, tangible accomplishments—pounds of apples picked, bushels of wheat harvested, etc. The same holds for pre-industrial manufacturing, where it was possible to pay for piece work.

Inside a large 20th century factory, the output of the *individual* worker was not as readily measured. Total output was still quantifiable, but one could not attribute a ton of steel or a shipment of refrigerators to an individual worker. Factory workers are paid to be in a particular place at a particular time. They clock in, clock out, and get paid by the hour.

For podcasts on the decline in manufacturing jobs and measuring output, see **Adam Davidson on Manufacturing** and **Kelly on the Future, Productivity, and the Quality of Life**. For a podcast illustrating a surprising data source in a high-tech industry, see **Varoufakis on Valve, Spontaneous Order, and the European Crisis**.

In the decades following World War II, as the share of employment on the factory floor declined, what increased was clerical work. Secretaries, sales clerks, and telephone operators, like factory workers, are paid to be in a particular place at a particular time. They might receive a biweekly salary, but they are expected to work set hours.

With the advent of the Internet, work is no longer necessarily fixed in time and space. Someone can answer email on a smart phone at any time, in any location. A friend of mine in sales and support for an educational software company is typical. Because his clients span the globe, he can be called at any time of day or night. Neither they nor his boss know where he is when he picks up the phone. Usually, he is in his basement office, but sometimes he is out of town at a conference or on a family vacation. As more workers hold jobs whose contours are shaped by the Internet, the once well-defined concept of "hours worked" becomes much less precise.

It is not just the individual worker for whom output and hours worked have become harder to pin down. Entire industries, such as education, health care, finance, and government are now increasingly prominent. In each of these sectors, the very definition of "output" is not clear.

Moreover, it is often the case that the value of these services is contingent on institutional arrangements. For example, the value to an individual of a particular educational credential will depend on the licensing requirements within the field in which the person is pursuing employment. The value of pooling mortgages into securities depends on the institutional role of rating agencies, government policies, and other highly contingent social arrangements.

Garett Jones has suggested (via Twitter) that "Workers mostly build organizational capital, not final output."[2] Building organizational capital means developing processes and capabilities within a firm. That is, in an institutions-intensive economy, many jobs consist of attempting to improve institutions. Business process improvement, standards-setting, and negotiating arrangements with other firms all fit within this framework.

Although we think of companies like Amazon, Apple, and Google as technology companies, these firms are institutions-intensive. Google's mechanism for auctioning ads has been the key to its profitability. Similarly, Apple's success depends in large part on institutional arrangements, such as its iTunes service and its App store. Amazon's encouragement and use of customer reviews has played an important

role in the success of its online sales efforts.

In neoclassical economics, the earnings of individuals are thought to be determined by individual characteristics and decisions. Economists write down models in which earnings accrue in a deterministic fashion to human capital and physical capital. However, these models are not realistic in an institutions-intensive economy.

One example of the complexity of today's economy can be found in a recent best-selling book, *The Immortal Life of Henrietta Lacks*, by Rebecca Skloot.[3] Cells taken from a tumor of Henrietta Lacks have been a tremendous boon to medical research, including help with the cure for polio. Skloot's book raises the issue of whether Lacks and her family were adequately compensated for this.

On the one hand, one could argue that the cells of a person's body are certainly the property of that individual, so that Lacks and her descendants are entitled to great wealth. After all, is not "self-ownership" a core foundation for the definition of property? On the other hand, one can argue that the value was created not by Lacks herself, but by the researchers.

Similar issues arise with regard to data on the Internet. Should Amazon be compensating the people who write helpful reviews? Should Facebook and Google be compensating people for providing information and services that those companies use to earn advertising revenue?

The technologically-determined income shares of neoclassical economic models do not describe our contemporary world. In an institutions-intensive economy, the pristine modeling of non-institutional factors misses out on the institutional contingencies that actually drive economic outcomes.

Footnotes

1. Ronald Coase, "Saving Economics from the Economists." *Harvard Business Review*, December 2012. (Available online at **http://hbr.org/2012/12/saving-economics-from-the-economists**.)
2. Tyler Cowen, **"The Wisdom of Garett Jones."** Marginal Revolution. Originally published on November 5, 2009.
3. Rebecca Skloot, *The Immortal Life of Henrietta Lacks.* New York (Random House, Inc.), 2011.

---

*Arnold Kling has a Ph.D. in economics from the Massachusetts Institute of Technology. He is the author of five books, including *Crisis of Abundance: Rethinking How We Pay for Health Care*; *Invisible Wealth: The Hidden Story of How Markets Work*; and *Unchecked and Unbalanced: How the Discrepancy Between Knowledge and Power Caused the Financial Crisis and Threatens Democracy*. He contributed to EconLog from January 2003 through August 2012.

For more articles by Arnold Kling, see the **Archive**.

▲ Return to top

(amagi), or "liberty." It is taken from a clay document written about 2300 B.C. in the Sumerian city-state of Lagash.

JOURNAL OF LAW, ECONOMICS & POLICY

VOLUME 9 FALL 2012 NUMBER 1

JOURNAL OF LAW, ECONOMICS & POLICY

VOLUME 9 FALL 2012 NUMBER 1

JOURNAL OF LAW, ECONOMICS & POLICY

VOLUME 9 FALL 2012 NUMBER 1

CONTENTS

ARTICLES

COMMENTS

DESIGN, INSTITUTIONS, AND THE EVOLUTION OF PLATFORMS

Richard N. Langlois[*]

ABSTRACT

Baldwin and Woodward[1] 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 Ferguson[2] 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 Farell & 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. Leibowitz & Stephen E. Margolis, *Path Dependence, Lock-In, and History*, 11 J.L. ECON. & ORG. 205 (1995); S.J. Leibowitz & 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 Grove[28] 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]

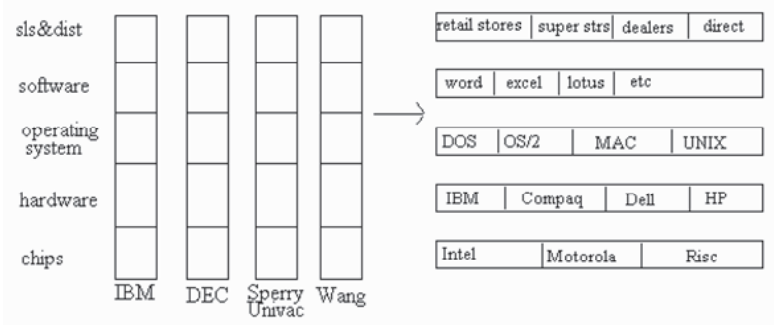


Figure 1.

Modular cluster in the computer industry, circa 1995.

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 Smith[35] 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 century[47]—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—forwent 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 Law[56]—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 firms[61]—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 hierarch, 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 hierarch, or the hierarch's commitment to delegate to representative institutions, to commit to an *ex post* allocation of returns that induces user reliance. The use of mediating hierarch 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. smehra@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* contracts[7] or *ex post* legal remedies[8] 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" model[9]—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 systems[11] 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/liberalisationandcompetitioninterventioninregulated sectors/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—itself. 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 regulators[20]—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).

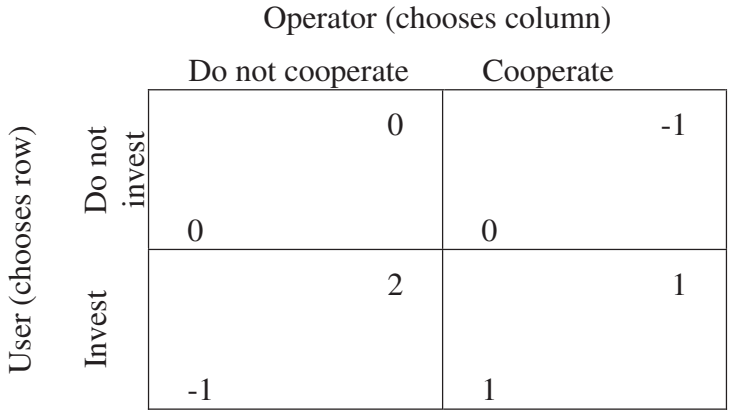
30 *See* Thomas Hazlett, David Teece & Leonard Waverman, *Walled Garden Rivalry: The Creation of Mobile Network Ecosystems* 5-16 (George Mason Law & Economics Research, Working Paper No. 11-50, 2011), *available at* http://ssrn.com/abstract=1963427.

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.



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' innovation[37]—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 antitrust 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 WEBLOGS AND SOCIAL MEDIA 27, 29 (2008), *available at* http://www.aaai.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 "privateering," 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 NPEs[29] 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 offices[39]—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; s*ee 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://fosspatents.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/nov06/11-02MSNovellPR.mspx.

56 Asay, *supra* note 52.

57 Jon Brodkin, *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 Triedman, *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*, MOCONEWS (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 filing[95]

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 smartphones

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 1:2011cv00654	Patent infringement suit on 3G phones, USB sticks, mobile hotspots and tablets
VirnetX	Apple	November 2011	Eastern District of Texas, 6:2011cv00563	Apple iPads, iPods, and iPhones alleged to infringe VirnetX patents
Intellectual Ventures	AT&T, T-Mobile, and Sprint Nextel	February 2012	District of Delaware, 1:2012cv00193	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),* MARKETSANDMARKETS (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 laws[104] 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 value[2] 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 is[10]—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 1988[14] 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 Bhagat, *Costs of Inefficient Bargaining*]; Sanjai Bhagat et. al., *The Shareholder Wealth Implications of Corporate Lawsuits*, 27 FIN. MGMT. 5, 24-25 (1998) [hereinafter Bhagat, *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 countersuits, 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 Lanjouw & Schankerman in 2001. *See* Jean O. Lanjouw & 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 methodology[22] 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) $r_t = \alpha + \beta r_t^m + \epsilon_t,$

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

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 Salinger[28] 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.

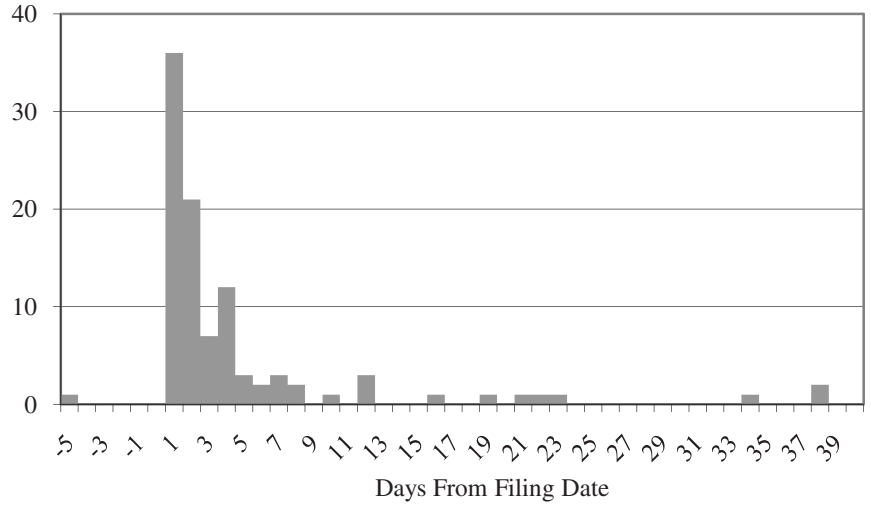
On the other hand, our estimates may be understated for another reason: investors may not receive news of the lawsuit within an event window around a filing date. With an announcement in a newspaper or major newswire, we can be reasonably sure that investors hear the news of the lawsuit within a day 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)
Patentee Litigant (Plaintiff)			
mean	-0.3%	-0.1%	-0.31%
median	0.0%	0.9%	
no. of observations	86	86	
Alleged Infringer (Defendant)			
mean	-2.6%	-1.8%	-1.50%
median	-1.4%	-1.9%	
no. of observations	82	82	
Combined (matched parties)			
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					
Patentee Litigants					
Base	-0.38%	(0.30%)	0.00%	-1.51	667
Definite infringement suits	-0.63%	(0.37%)*	-0.45%	-2.18*	412
Alleged Infringers					
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

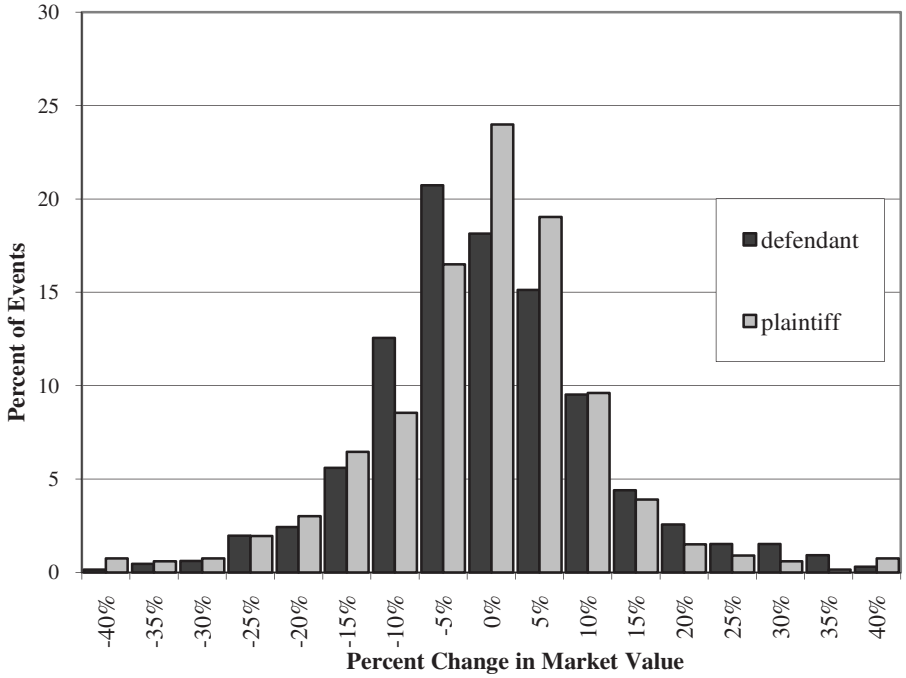
who have less at stake. Furthermore, costs may be shared among multiple defendants, reducing the individual firm costs.

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 subgroups. 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

Sample: Matched Parties

	Alleged Infringer	Patentee Litigant
Firm characteristic		
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 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
Patentee Litigant			
Summary Judgment	.95	.40	8
Verdict	1.04	.78	51
Alleged Infringer			
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

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

---

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 ith firm's CAR, and x is the ith 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%
1996–99			
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%

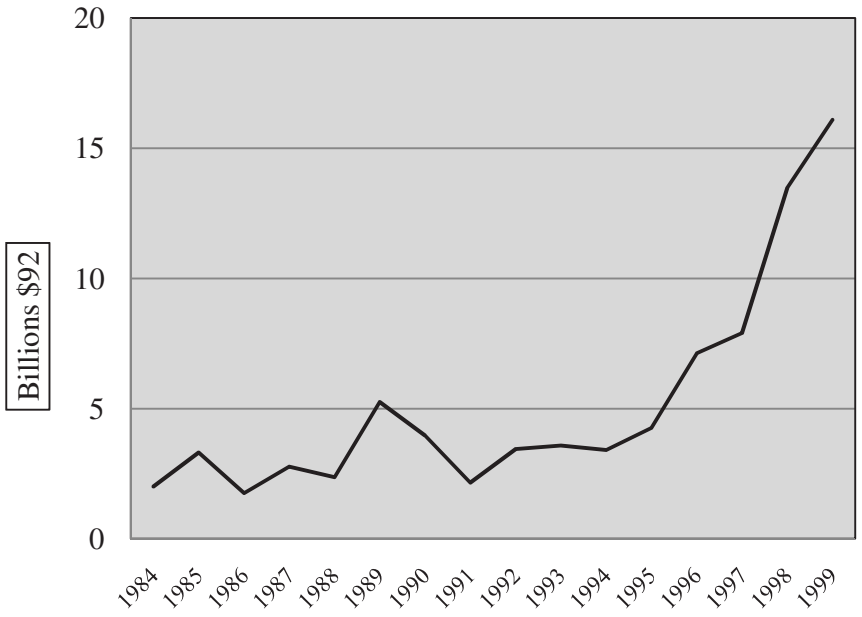
Note: The annual cost of litigation is the mean CAR times the market capitalization of each firm's common stock divided by a GDP deflator and by the aggregate Tobin's *Q* (market value divided by replacement value of capital including R&D). Firm infringement risk is the expected annual cost of litigation. 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 1991was 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.ssrn.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
Plaintiff/patentee litigant				
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)
Defendant/alleged infringer				
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

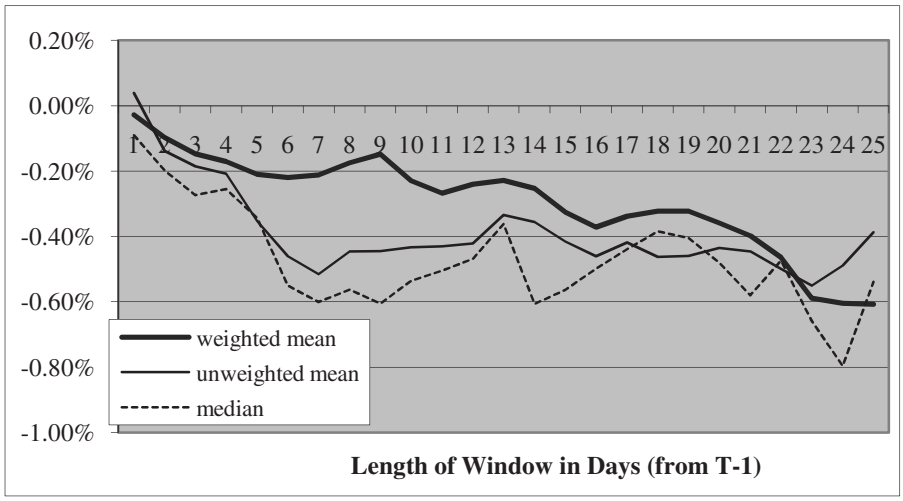
Note: Probit regressions. Robust standard errors in parentheses. Bold estimates are significant at the 5% level or better. Regressions include industry dummies (not shown).

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); s*ee 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 hierarchal 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.mttlr.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 convictions[16] 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 deportation[23] and bars to reentry[24] 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-BvQfKcf44Zb8psJJtNU_HD8aP3o5hP44Zni6EhdzmWlF8F4pcTDielodSh54tPvHtRGDX59stfewjp5TsUDciHUfKeXedtKm2J0TnrN&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 counsel[53]—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_guiltypleas_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 them[68] 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 cost[83] 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).

piled[91]—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 courts[97] 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&surl=/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 individuals[106] 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 difficultly 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 services[136] 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 money[138] 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 then [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.

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-ldquoriskrdquo-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 places[171] 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 others[191] without imposing criminal sanctions,[192] existing punitive and prohibitionary 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 2011[194] 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 food[1] 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_.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 distances[38] 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 nontraditional 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.*

USDA[70] 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=itrnotallissues&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. § 350l(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 rulemaking, 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 safer[114]—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 producers[145]—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 tasks[157] 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, atA531 ("'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, atA531.

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 Defense[175] 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 local[179]—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(l), 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.

December 31, 2012

Natural Gas: From Shortages to Abundance in the U.S.

Paul L. Joskow[1]

The recent dramatic and largely unanticipated growth in the current and expected future production of shale gas, and the related developments in the production of shale oil, have dramatically changed the energy future of the U.S. and potentially of the world compared to what experts were forecasting only a few years ago. These changes would not have been realized as quickly and efficiently absent deregulation of the wellhead price of natural gas, unbundling of gas supplies from pipeline transportation services, the associated development of efficient liquid markets for natural gas, and reforms to the licensing and regulation of prices for gas pipelines charge to move gas from where it is produced to where it is consumed. This economic platform supported the integration of technological advances in vertical drilling, down-hole telemetry, horizontal drilling, monitoring and control of deep drilling equipment, and hydraulic fracturing to exploit economically shale gas deposits that were identified long ago, but considered to be uneconomical until recently.

I. Natural Gas Wellhead Price and Pipeline Regulation

Federal regulation of the natural gas industry began with the Natural Gas Act of 1938 (NGA). The NGA gave the Federal Power Commission (FPC), later the Federal Energy Regulatory Commission (FERC), the authority to license the construction and expansion of new interstate natural gas pipelines, to ensure that they are operated safely, and to regulate the prices

[1] Alfred P. Sloan Foundation, 630 Fifth Avenue, Suite 2550, New York, NY 10111 (email:joskow@sloan.org) and Massachusetts Institute of Technology. The Alfred P. Sloan Foundation has made several grants to support research related to shale gas resources, environmental impacts and regulation, and political economy. The author is on the Boards of Directors of Exelon Corporation and TransCanada Corporation. The views expressed here are those of the author and do not reflect the views of MIT, the Alfred P. Sloan Foundation, Exelon, TransCanada, or the Sloan Foundation's grantees.

that these pipelines charge for supplying gas to local distribution companies (LDC) and large direct service consumers (e.g. electric utility plants). Until the 1990s, pipelines sold a "bundled" product consisting of the natural gas they purchased from gas producers combined with the pipeline services required to transport the gas to their customers. Gas production, gas transportation, and gas distribution were linked together by long-term contracts. However, the contract prices that the interstate pipelines paid for natural gas in producing areas were initially not subject to regulation. This situation changed in 1954 when the Supreme Court determined in the *Phillips* decision (347 U.S. 672 (1954)) that under the NGA the FPC had the authority to regulate the wellhead prices of natural gas purchased by interstate pipelines for resale (Breyer and MacAvoy 1973).

The FPC then proceeded to try to exercise that authority. The basic theory that guided FPC regulation of natural gas field prices started with the assumption that the supply function for natural gas was inelastic in the short run, was upward sloping in the long run, and that the demand for natural gas would increase significantly as the pipeline transportation infrastructure expanded. Absent regulation of natural gas prices, low-cost producers would earn economic rents in the long run as the demand for natural gas increased and unregulated market prices increased along with it. The FPC tried to use cost of service principles to capture the rents associated with low-cost gas wells for consumers by keeping the prices low-cost producers were allowed to charge low while setting higher prices for higher-cost wells to provide incentives for producers to expand supplies sufficiently to balance supply and demand. This multi-price regulatory system was then "harmonized" by requiring pipelines to average together the prices they paid for gas from wells with different regulated prices and to charge the average price to their customers.

Breyer and MacAvoy (1973) and MacAvoy (2000, pages 12-15) describe the almost comical efforts of the FPC to apply cost of service regulatory principles developed for regulating, say, a single local monopoly gas distribution company, to thousands of natural gas wells, many producing both unregulated crude oil and regulated natural gas, to many production districts, in the context of an intrastate market where wellhead prices were unregulated. As described nicely by Breyer and MacAvoy and by Breyer (1982, Chapter 13), these regulatory efforts, like most rent control regulations, succeeded in keeping field prices below competitive market levels to the benefit of consumers with access to that gas, but eventually led to growing shortages of natural gas in the interstate market. Domestic natural gas production hit its local peak in 1970 (not reached again until 2010 as shale gas production ramped up) and shortages began to appear in the late 1960's and were growing before the first oil shock in late 1973.

Thus, at the time of the first oil shock in 1973-74, natural gas wellhead price regulation had already been in place for 20 years and was already causing significant gas supply shortages. To deal with these shortages a complex rationing scheme to allocate scarce supplies among existing customers was developed and potential new customers were turned away. The shortages continued to grow during the 1970s as oil prices increased, leading to an increase in the demand for natural gas since petroleum and natural gas were substitutes on the margin (e.g. as a boiler fuel for electric generating plants and manufacturing and in certain process industries). Breyer (1982, 244-245) estimates that curtailments of firm contracts for natural gas grew by a factor of 20 between 1970 and 1976 and that some states experienced very significant reductions in deliveries to industrial consumers. MacAvoy and Pindyck (1973, 491) predicted growing shortages from regulation out to 1980, the end of their simulation period, but the rapid end of shortages if wellhead prices were deregulated.

After creating the shortages, Congress and the FPC then spent the next fifteen years trying to adopt and implement policies to mitigate the shortages while only slowly phasing out the rent transfers associated with the existing wellhead price regulatory system. The basic regulatory mechanism to do so was to set different prices for different "categories" of natural gas and then to require pipelines to charge customers the average cost of their contracts with producers; high prices for various categories of "new" and "high cost" gas and much lower prices for various categories of low-cost "old gas" were averaged together to produce a sale price between the low and the high authorized wellhead ceiling prices.

In November 1978 Congress passed the Natural Gas Policy Act of 1978 (NGPA) to implement this strategy. According to MacAvoy (2000,15), eventually 30 different categories of natural gas with widely varying ceiling prices were created under the NGPA. In response to this new regulatory framework, pipelines rushed to sign long-term contracts with suppliers that were now willing to sell additional supplies of "new gas" at high regulated or unregulated import prices. It was expected that the NGPA would lead to rising average prices and ultimately to de facto deregulation of wellhead prices as "old gas" supplies declined, and existing "old gas" wells were able to be reclassified as high cost producers subject to much higher price ceilings, and supplies of high-priced "new gas" became a larger and larger share of the "blend." The NGPA did not work out as anticipated.

II. Regulatory Induced Surpluses Lead to Pipeline Restructuring

The second oil shock(s) of 1979-1980, when real crude oil prices quickly doubled, occurred soon after the NGPA was passed. The jump in petroleum prices exacerbated shortages of natural gas as customers sought to switch from oil to natural gas. Pipelines responded to the existing and apparently growing shortage of natural gas by taking advantage of the opportunities

created by the NGPA to negotiate long term "take-or-pay" contract supply commitments with producers for "new gas" at high ceiling prices and for imported gas at high unregulated prices. These contracts committed them to take minimum quantities of gas based on the pricing provisions in the contracts. Pipelines expected to be able to sell the gas to consumers because the prices of all these categories of natural gas were blended together, creating an average sale price significantly below the ceiling prices established by the NGPA (MacAvoy 2000,15) and the demand for natural gas at these prices was expected to continue to grow. However, as blended gas prices rose, the demand for natural gas fell. This effect was exacerbated by falling oil prices after 1981-- real crude oil prices peaked in 1981 and fell by nearly 2/3 by 1986 making gas less economical compared to oil at the now higher regulated bundled price charged by pipelines. This in turn led to a "take-or-pay contract bubble" in which the demand for natural gas by LDCs and direct service customers was significantly lower than the contractual obligations undertaken by the pipelines. (See Makholm 2012 for an excellent history of the evolution of natural gas and oil pipeline regulation.)

Efforts to deal with the "take or pay contract surplus" led to a series of additional regulatory policy changes that dramatically changed the structure of the natural gas pipeline sector. In the early 1980s, as price sensitive gas consumers switched from gas to oil and coal, some pipelines created Special Marketing Programs (SMPs) to retain these customers. SMPs allowed selected (elastic) customers to purchase "unbundled" transportation services from pipelines, so that these customers could buy gas directly at lower prices in producing areas and simply pay the pipelines to transport it, bypassing the higher regulated prices. However, in 1985 the DC Circuit declared that these arrangements were discriminatory and ordered FERC to terminate them. FERC responded with Order 436 which allowed pipelines voluntarily to offer

unbundled pipeline services as long as they offered these services on a non-discriminatory basis to all customers. Eventually all major pipelines voluntarily agreed to offer unbundled transportation services to all customers.

Of course, this exacerbated the "take or pay contract problem" as a growing number of customers sought unbundled transportation service to "bypass" the high regulated contract prices reflected in bundled pipeline charges. The Court of Appeals for the D.C. Circuit upheld most of the provisions of order 436, but remanded the challenges to it to FERC to resolve the "take-or-pay" problem. In 1987 FERC issued order No. 500 which encouraged pipelines to negotiate buy-outs of their contractual obligations to producers and allowed them to pass along a portion of these costs in pipeline charges to their customers (a classic regulatory solution to a stranded cost problem.) and in 1989 Congress passed the Natural Gas Wellhead Decontrol Act of 1989 (NGWDA) which accelerated the process for deregulating wellhead prices and integrated intra-state and interstate markets. By January 1, 1993 all wellhead gas price regulation came to an end, though little gas was still subject to field price regulation by 1990 or so.

Faced with a natural gas market that was now largely deregulated, the challenge of dealing with the large overhang of high-priced take or pay contracts, and court orders requiring FERC to deal with these issues in a non discriminatory fashion, in 1992 FERC then issued order 636 which made unbundling of pipeline services mandatory, permitted pipelines to market natural gas only through "ring fenced" affiliates, required pipelines to provide a number of additional unbundled pipeline services (e.g. storage, capacity release programs) and to provide information about available pipeline services and prices to increase the flexibility with which LDCs and large customers could use pipeline capacity. As this new structure evolved, FERC also adopted more and more forms of "light handed" regulation of pipeline transportation rates,

relying increasingly on competition between existing and new pipeline developers to negotiate service contracts with gas producers through competitive bidding processes.

III. Natural Gas Markets Mature During the 1990s

By the early 1990s, wellhead price regulation had come to an end, the intra-state and interstate markets had been integrated, the natural gas production sector was governed by competitive market forces, and gas shortages (and federal restrictions on the use of natural gas by power plants) disappeared. The natural gas market matured during the 1990s as liquid gas trading hubs (e.g. Henry Hub and Dawn Hub) developed, liquid spot, term, and derivatives markets developed, geographic basis differences declined. Wellhead prices were low and relatively stable during most of the 1990s. Domestic production increased slowly between 1991 and 2000 (+6%) while consumption grew more quickly. The difference was made up by an increase in imports (+15%), mostly from Canada.

The natural gas pipeline sector had also been substantially restructured to support both a competitive natural gas market and the entry and expansion of competing pipelines. Light handed regulation of pipeline certification and of unbundled transportation prices based on open season auctions and flexible bilateral transportation contracts replaced rigid cost-of-service regulation of pipeline services, except for captive customers who did not have access to competing pipelines. Pipeline capacity expanded and exploration, development and the geographic distribution of natural gas production shifted slowly to new (and more expensive) on-shore areas and off-shore producing areas.

IV. The Gas Surplus Comes to an End

Natural gas prices remained low and stable during the 1990s and many expected this situation to persist for many years. However, natural gas prices began to rise during 2000 and

both increased significantly and became much more volatile over the next few years. The received wisdom (ex post) in the natural gas industry and within the federal government at that time was that there had been a gas supply "overhang" during the 1990s and that as demand caught up with supply more expensive gas production sources would have to be relied upon to balance supply and demand (National Petroleum Council 2003, AEO 2003, AEO 2004), including more imports from Canada, the construction of a pipeline to bring new supplies of gas from Northern Alaska through Canada to the U.S., and a large expansion of LNG imports from the rest of the world (2003 AEO, 41-42, 2004 AEO, 43-44).

The recent dramatic growth in shale gas production was unanticipated early in this century. The reference case in the Energy Information Administration's (EIA) 2003 Annual Energy Outlook (AEO) did envision growing domestic production "unconventional" sources (coal bed methane, tight gas formations, and shale deposits, 2003 AEO, 35), but shale gas gets barely more than this mention and the forecast future production of shale gas identified is quite small. The 2004 AEO reflects a similar perspective. Unconventional gas production is forecast to account for 43% of lower 48 gas production, though the contribution of shale gas was forecast to be quite small (2004 AEO, 36, 38). LNG imports were forecast to be a larger incremental source of supply than all unconventional gas resources combined and a long list of proposed LNG import facilities is highlighted (2004 AEO, 41). The National Petroleum Council's (2003) report on natural gas did not even mention shale gas, but focused on opening up more off-shore and on shore areas to exploration and development, the construction of a pipeline to bring gas from Northern Alaska to the lower-48 states, and speedier certification of LNG import facilities.

This view of the future supplies, demand and prices for natural gas began to change only a few years ago. Much of the growth of U.S. natural gas production is now forecast to come

from domestic shale gas deposits. Shale gas production is forecast in the 2012 AEO to increases to 49% of domestic gas supplies by 2035 in the reference case and total unconventional domestic gas production and total non-conventional gas supplies account for 70% of domestic supplies (2012 AEO, 93). By 2010 shale gas production was already 5 trillion cubic feet and forecast to rise to 13.6 trillion cubic feet in 2035 (2012 AEO Early Release, 1) and more recent forecasts of future shale gas production continue to grow (2013 AEO Early Release, 1). (Oil from deep shale deposits and tight sands has and is also forecast to increase dramatically, but that is a subject for another paper.) The 2012 AEO forecasts the U.S. to become a net exporter of LNG by 2016 and an overall (including pipelines) net exporter of natural gas by 2021 rather than a large net importer of natural gas. The 2013 AEO (Early Release) advances these dates.

V. The Shale Gas Revolution

What happened to change so dramatically the supply and import picture for natural gas (and oil) in such a relatively short period of time? The answer is unanticipated increases in production of natural gas from shale deposits located 1,000 to 13,500 feet below the surface (MIT, page 40). Shale gas is natural gas that has been trapped in deep shale deposits, which are fine-grained sedimentary rocks with very low permeability. Shale gas deposits differ from conventional gas deposits which are composed of gas that has migrated to pools in permeable rocks typically located much closer to the surface. So-called "shale plays" are located in many areas of the U.S.; in Texas, the South, the Northeast, the mid-West, the Rocky Mountain region, and in Western Canada (as well as in many other countries). In 2004 there was very little gas produced from shale deposits. Today shale gas already accounts for about 25% of domestic production and has helped to keep natural gas prices at relatively low levels for the last few years. In 2004 monthly wellhead prices varied between $5.73/MCF and $10.33/MCF, while in

during the first ten months of 2012 monthly wellhead prices varied between $1.89/Mcf and $3.50/Mcf, although the longer run equilibrium price is almost certainly significantly higher than these recent levels (2012 AEO, 92). How much higher gas prices will go is the subject of considerable debate at the moment, though the latest EIA forecasts (2013 AEO Early Release) have a lower gas price trajectory than the 2012 forecasts.

It has been known for many years that natural gas was embedded in deposits of shale deep in the earth. Unlike conventional deposits of natural gas, this gas did not accumulate in relatively shallow pools in permeable rock formations, but rather was embedded in deep shale deposits with low permeability. Commercial production of this gas was thought to be quite uneconomical with then existing technology. Several technological advances came together in the late 1990s to make the development of shale gas economical: advances in deep vertical drilling technology, horizontal drilling technology, down-hole telemetry, monitoring and control of drilling equipment, and hydraulic fracturing in horizontal wells.

In a typical shale gas well, a deep vertical well is drilled toward the shale deposit with drilling equipment and drilling "mud" to cool and lubricate the drill bit and to bring drilling fragments to the surface. The drilling equipment is withdrawn and the vertical well is lined with cement and steel. The steel and cement lining is reinforced extensively close to the surface and where the vertical well penetrates drinking water aquifers (typically far above the shale play). Drilling equipment is then reinserted, the vertical well is completed down to the shale deposit and then one, or typically multiple, horizontal wells are drilled for long distances along the shale seam. When the horizontal design distance is reached, the drilling equipment is withdrawn and the horizontal well is lined with steel and cement. A fracturing tool is inserted into the well which uses small explosive devices to create holes in the horizontal cement and pipe and to

"fracture" the shale deposit so that gas can flow more easily into the well. High pressure "fracking" fluid composed of water, sand and various chemicals (about 90% water, 9% sand, and 1% a variety of chemicals to keep the plumbing clear and gas flowing) is then inserted into the horizontal portion of the well to further loosen and keep open the shale rock formation so larger quantities gas can flow more easily. A typical shale gas well takes a few weeks to complete, though the hydraulic fracturing process itself takes a few days. http://www.energyfromshale.org/shale-extraction-process (accessed May 20, 2012)

Although most of the public controversy about shale gas focuses on hydraulic fracturing, it is not a new technology. Indeed, hydraulic fracturing has been used to stimulate well flows from traditional vertical oil and gas pools since at least 1947 and probably earlier. Public-private partnerships engaged in research, development, and demonstration projects involving hydraulic fracturing, horizontal drilling, imaging, and telemetry from the 1970s through the 1990s. The first economical use integrating all of these technological developments is generally attributed to the efforts of a Mitchell Energy project in the Barnett shale play near Fort Worth in 1998 (http://www.economist.com/node/21558459 accessed November 22, 2011). The first major applications were slowly introduced in the Barnett shale play around Fort Worth, Texas in the early years of this century. The number of horizontal wells operating in the Barnett shale area increased from 400 in 2004 to 10,000 in 2010 and shale gas production increased from almost nothing in 2000 to about 5 billion cubic feet per day today. Information gained from the experience in the Barnett shale has helped to increase the efficiency of the exploitation of more recent shale gas production areas. Indeed, experience with shale gas production appears to have led to dramatic increases in productivity.

VI. The Supporting Role of Reforms of Wellhead Pricing and Pipeline Regulation

Although the recent extraordinary developments in shale gas production are primarily a story about the interaction between geology, natural resources, the development of new technologies for using these resources, and associated learning and continuous improvement, the deregulation and regulatory reform policies of the 1980s and 1990s certainly played a role in supporting these developments. If wellhead prices were still regulated as they were from 1954 to at least 1980, the administrative apparatus for establishing prices alone would have slowed down development of these new resources. The cost-of-service mentality of the federal regulators would have undermined incentives to invest in a new and changing production technology. The absence of liquid spot markets, short-term term contract markets, derivative markets along with pipelines and customers tied up for years with long term contractual commitments would have reduced market opportunities for shale gas, especially in the early years when many of the most innovative producers were small firms.

Perhaps more importantly, many of the most promising shale plays (e.g. Marcellus, Utica) are not located in traditional producing areas and do not have in place the pipeline infrastructure to get the gas to market. The development of faster and more efficient pipeline permitting rules, pricing and contracting flexibility, and the evolution of a merchant pipeline sector that can develop and build new pipeline projects quickly and take on merchant financial risk, have made it possible for new and existing pipeline investors relatively quickly to build new pipeline capacity to serve these new producing areas.

VII. Benefits to the U.S.

While the title of my colleague John Deutch's *Wall Street Journal* (2012) Opinion piece "The U.S. Natural-Gas Boom Will Transform the World," may be a little over the top, it is clear that there are substantial economic and national security benefits from a more abundant North

American supply of less costly natural gas (Deutch 2011). Less costly natural gas benefits residential, commercial and industrial consumers through lower prices. The development of domestic natural gas deposits, rather than deposits in other countries that would have been transported to the U.S. as LNG, has created an economic boom in the areas where these developments are taking place, and has a positive, though necessarily uncertain, U.S. employment effect in the natural gas (and shale oil) sectors, associated infrastructure sectors, and the sectors that manufacture the equipment used in the oil and gas and pipeline sectors (IHS CERA 2012). Cheap natural gas is in turn leading to higher capacity factors for existing natural gas electric generating plants, reducing the dispatch of coal plants, reducing CO2 emissions as gas-fired power plants emit as much as 50% less CO2 per unit of electricity than do coal plants. The return of relatively cheap natural gas is especially beneficial to certain industrial sectors, especially petrochemicals, fertilizer, cement, metallurgical process industries, and manufacturing plants that use natural gas as a boiler fuel (Ben Casselman and Russel Gold 2012, HIS CERA 2012). And, of course, it makes U.S. firms more competitive with firms in Asia and Europe where natural gas prices are two to four times higher than they are in the U.S.

There are also potential national security benefits. As recently as 2004, the conventional wisdom was that substantial imports of LNG from the Middle East and North Africa would have been required to balance the supply and demand. As with petroleum imports, the U.S. would have been at risk for natural or political disruptions in LNG trade and would have been transferring significant wealth to natural gas producing countries. The U.S. is now forecast to be a modest net exporter of natural gas by 2035, though there is considerable uncertainty about how important LNG exports, as opposed to pipeline exports to Mexico, will be. This in turn has some implications for long run equilibrium prices that I will not explore here. I note only that

liquefaction and shipping of LNG is expensive so that there will continue to be a significant basis difference in gas prices between the U.S., Asia and Europe. (International Energy Agency 2012, 130, EIA 2012 and NERA 2012).

VIII. Environmental Issues

Of course, environmental regulation of natural gas exploration, development, and production continues at the state and federal levels. While the shale gas industry has a good environmental record and there are relatively few cases that have been credibly identified of serious environmental impacts (MIT 2011), shale gas development has attracted substantial public controversy in some areas of the country that has and will continue to slow down development of this resource if the public cannot be convinced that shale gas can be developed safely and responsibly. Over the next decade hundreds of thousands of new shale gas wells are expected to be drilled across the United States. Drilling likely will take place in areas of the country which do not have much recent experience with oil and gas drilling and its regulation (e.g. Pennsylvania, New York, Ohio), where new or revised regulatory frameworks need to be developed and strengthened, and where the public is more concerned about the environmental impacts of shale gas development than in states that have a long history of significant oil and gas production.

There are numerous pathways through which the various components of the exploration, development and production process can have potentially adverse environmental effects (Resources for the Future (a) 2012). The scope and quality of state regulation varies widely (Resources for the Future (b) 2012), best practices need to be developed, the division of responsibility between state and federal regulators needs to be defined more clearly, the societal

costs of various environmental impacts need to be measured much more precisely, and the industry needs to be more transparent about its operations to gain public acceptance.

Most of the focus of public opposition to shale gas development has been on the relationship between "fracking" and contamination of groundwater with natural gas. The 2010 documentary *Gasland* left the indelible picture that fracking leads to large quantities of natural gas being released into drinking water, which in turn could be set on fire inside people's homes. While this kind of contamination is possible in theory, the widespread direct leakage of natural gas from hydraulic fracturing into aquifers used for drinking water is unlikely because the shale gas deposits where hydraulic fracturing takes place are typically located thousands of feet below these water aquifers. If releases of methane from shale gas development into groundwater take place they are most likely to come from failures in the casings of the vertical well bores, shallow conventional gas deposits that are disturbed during the vertical drilling process, or from dissolved methane in fracking fluids that are spilled on the surface.

There are few credible documented cases of shale gas development leading to significant releases of methane into groundwater, fires in houses when faucets are turned on, etc., though the necessary "before and after" data to perform good studies is lacking (Alan Krupnick, Lucija Muehlenbachs, and Karlis Muelenbachs, 2011). Natural seepage of methane from shallow conventional deposits, leaks from conventional drilling from shallow pools of gas, and water well drilling that hits shallows deposits of natural gas are likely more significant sources of methane in aquifers. (I am told that releases of natural gas into drinking water in areas around Fort Worth pre-date shale gas development) It is clear, however, that regulations governing well completion to ensure proper steel casing and cementing of the vertical portions of shale needs to be tightened to ensure that such leaks into drinking water aquifers do not occur. More good

research that examines water quality before and after well completions and studies the pathways that may characterize water contamination is certainly needed as well.

The publicity around the relationship between hydraulic fracturing and drinking water contamination with natural gas has led to under-appreciation for other pathways of potential environmental concern (U.S. DOE 2011). These include contamination of groundwater from improper storage and disposal of drilling mud and toxic chemicals in hydraulic fracturing fluids, some of which may be unnecessary (e.g. diesel oil) or can be replaced at low cost with less toxic chemicals, air emissions from on site electric generators and vehicles, and disruption of local communities from large increases in heavy truck traffic bringing equipment into drilling sites and shipping out the waste and used equipment associate with drilling and hydraulic fracturing. Hydraulic fracturing also uses prodigious volumes of water during the well completion process (a few days for each horizontal well). This raises concerns not only about the disposal of the water and the chemicals in it, but the effects on local water resources and their allocation to different uses. Releases of methane into the atmosphere have also been identified as a concern because methane is a very potent greenhouse gas. However, the studies on this subject yield results that are all over the place, and if this is a problem it must be resolved across the entire oil and gas production and pipeline industries.

The shale gas boom has proceeded more rapidly than has the development of the associated environmental regulation, public information, and public education about the real and imagined risks and the actions state and federal governments are taking to deal with the important risks. The U.S. DOE's Shale Gas Production Subcommittee's Report (2011) makes many good suggestions for reducing environmental impacts and improving transparency and

public education. Developing good environmental regulatory frameworks is essential if the potential for shale gas development in the U.S. is to be realized.

Of course there are also environmental benefits from greater supplies of cheap natural gas, especially as it relates to displacing coal from utility and industrial boilers. While increased use of natural gas in North America is not going to achieve the greenhouse gas reductions necessary to stabilize global temperature increases it will provide a medium term economic benefit and reduce GHG emissions compared to forecasts made only a few years ago.

REFERENCES

Breyer, Stephen. 1982. *Regulation and Its Reform*. Harvard University Press. Cambridge and London.

Breyer, Stephen and Paul W. MacAvoy. 1973. "The Natural Gas Shortage and the Regulation of Natural Gas Producers." *Harvard Law Review*, 86(6):941-987.

Casselman, Ben and Russell Gold. 2012. "Cheap Natural Gas Gives New Hope to the Rust Belt." *The Wall Street Journal*, October 24. http://online.wsj.com/article/SB10000872396390444549204578020602281237088.html -- accessed November 22, 2012

Deutch, John. 2011. "The Good News About Gas." *Foreign Affairs*. 90(1): pp. 82-93.

Deutch, John. 2012. "The U.S. Natural Gas Boom Will Transform the World." August 15: A13.

IHS CERA. 2012. *America's New Energy Future: The Unconventional Oil and Gas Revolution and the U.S. Economy*. October. http://www.ihs.com/products/cera/energy-report.aspx?ID=106592439 -- accessed November 22, 2012

International Energy Agency (IEA). 2012. *World Energy Outlook 2012*. Paris. http://www.worldenergyoutlook.org/ -- accessed November 22, 2012.

Krupnick, Alan, Lucija Muehlenbachs, and Karlis Muelenbachs. 2011. "Shale Gas and Groundwater Contamination: Thoughts on a Recent Study." Resources for the Future. http://www.rff.org/news/features/pages/shale-gas-and-groundwater-contamination.aspx -- accessed November 20, 2012.

MacAvoy, Paul W. 2000. *The Natural Gas Market: Sixty Years of Regulation and Deregulation.* Yale University Press. New Haven and London.

MacAvoy, Paul W. and Robert S. Pindyck. 1973. "Alternative Regulatory Policies for Dealing with the Natural Gas Shortage." *Bell Journal of Economics and Management Science*. 4(2): 454-498.

Makholm, Jeff G. 2012. *The Political Economy of Pipelines*, University of Chicago Press. Chicago and London.

Massachusetts Institute of Technology (MIT). 2011. *The Future of Natural Gas*. http://mitei.mit.edu/publications/reports-studies/future-natural-gas --- accessed November 1, 2012.

National Petroleum Council (NPC). 1973. *Balancing Natural Gas Policy: Fueling the Demand of A Growing Economy*. Volume I. Washington, D.C.

Naturalgas.Org, "The History of Regulation." http://www.naturalgas.org/regulation/history.asp --accessed November 20, 2012.

NERA Economic Consulting (NERA). 2012. "Macroeconomic Impacts of LNG Exports from the United States." http://www.fossil.energy.gov/programs/gasregulation/reports/nera_lng_report.pdf --- accessed December 20, 2012.

Resources for the Future (a). 2012. "Risk Matrix for Shale Gas Development." http://www.rff.org/centers/energy_economics_and_policy/Pages/Shale-Matrices.aspx -- accessed November 20, 2012.

Resources for the Future (b). 2012. "A Review of Shale Gas Regulations By State." http://www.rff.org/centers/energy_economics_and_policy/Pages/Shale_Maps.aspx --- accessed November 20, 2012.

The Economist. 2012. "An unconventional Bonanza." Special Report. July 14th. 3-18.

U.S. Department of Energy. 2011. *Shale Gas Production Subcommittee 90-Day Report*. Secretary of Energy Advisory Board. August. Washington, D.C. http://www.shalegas.energy.gov/resources/081811_90_day_report_final.pdf -- accessed November 20, 2012.

U.S. Energy Information Administration (AEO). Various Years. *Annual Energy Outlook (AEO)* (Early Release and Final). http://www.eia.gov/forecasts/aeo/ --- accessed November 1, 2012.

U.S. Energy Information Administration (EIA). 2012. "Effects of Increased Natural Gas Exports on Domestic Energy Markets." http://www.eia.gov/analysis/requests/fe/pdf/fe_lng.pdf --- accessed December 1, 2012,

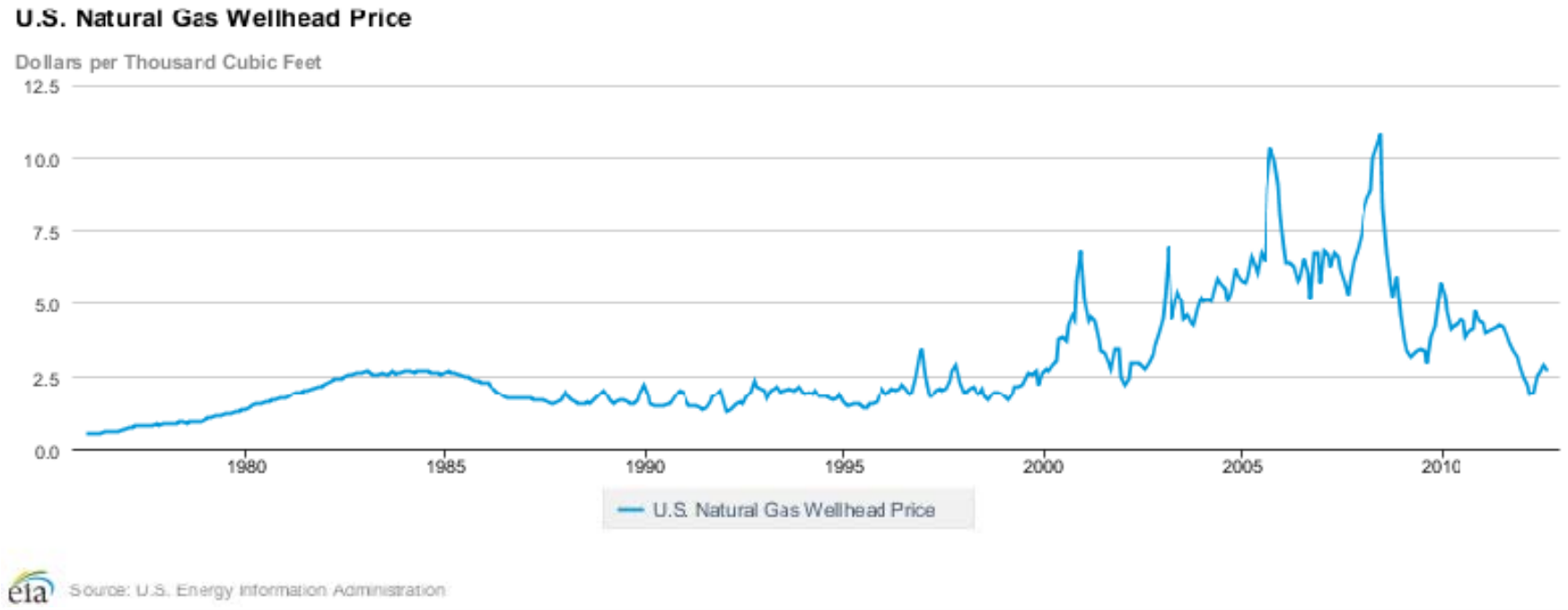
Natural Gas: From Shortages to Abundance in the U.S.

Paul L. Joskow

January 4, 2013

SUMMARY

- Tortured history of natural gas wellhead price and interstate gas pipeline regulation, deregulation and regulatory reform 1938- 1993: creating and responding to natural gas shortages and then contract surpluses 1970-1990
- Deregulation of wellhead prices, unbundling, and pipeline regulatory reform ultimately created a platform conducive to efficient static and dynamic performance on the supply and demand sides of the natural gas market
- Shale gas resources became economical to exploit due to technological innovations in drilling technologies that could easily be exploited without the burdens of inefficient economic regulation
- The magnitude of shale gas production was a "surprise". Early 2000's expectation was that U.S. would become a large importer of natural gas from Canada, Alaska, LNG from Middle East and North Africa.
- Expected North American production of shale gas keeps increasing year after year, gas prices have fallen significantly, but are probably below long run equilibrium levels, and the U.S. is now expected to be a gas exporter by about 2020.
- Natural gas consumers will benefit from lower gas prices, the trajectory of CO2 emissions will be lower, and there are national security benefits from going from an expected gas importer to an expected gas exporter
- Unresolved environmental concerns, along with misleading information and a lack of transparency, have led to public opposition in some areas. These issues can be and must be resolved if shale gas is to reach its full potential



Source: U.S. EIA http://www.eia.gov/dnav/ng/hist/n9190us3m.htm 12/26/12

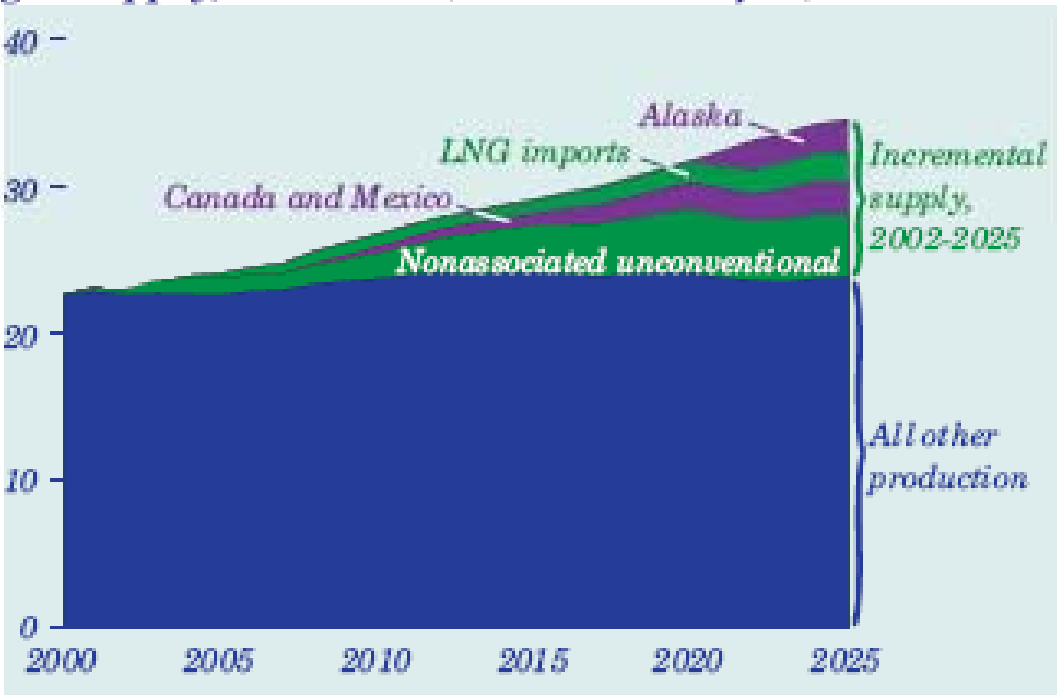
Natural Gas Regulation, Deregulation and Regulatory Reform

- 1938 --- FPC pipeline regulation of bundled gas and pipeline services
- 1954 --- *Phillips Decision*: FPC Regulation of wellhead prices begins
- 1954 – 1979 --- Various approaches to cost of service regulation applied to set maximum wellhead prices.
- Blending and bundling of gas contracts by pipelines key to implementation
- 1970 – 1981 --- Shortages and rationing
- 1978: Natural Gas Policy Act: old gas, new gas, high cost gas ; category by category ceiling prices.
- Pipelines rush to sign long term contracts for “new gas”with high ceiling prices and “blend” diverse contract prices together to create average cost-based sale prices
- 1981: Oil prices peak and then fall 2/3 by 1986. “Take or pay” contract surpluses.

Natural Gas Regulation, Deregulation and Regulatory Reform

- 1981- 85: SMPs for customers with fuel switching capabilities and the start of unbundling. Clear price discrimination leads to rejection by the courts
- 1985: FERC Order 436 voluntary unbundling available to all customers
- 1987: FERC Order 500 and take or pay contract renegotiation
- 1989: Natural Gas Wellhead Decontrol Act
- 1992: FERC Order 636 makes unbundling mandatory
- 1993: Regulation of wellhead prices formally ends
- 1990s: Competitive natural gas markets and light handed regulation of pipeline services mature --- prices low and stable
- 2000: (ex post) "Gas bubble" ends, gas prices rise and become volatile.
- 2001-2006: Large increase in pipeline imports from Alaska, Canada and LNG imports from other countries and higher prices expected to balance U.S. supply and demand in the future
- 2007: Shale gas slowly begins to be recognized as important future supply source
- 2009: Shale gas production starts to have significant effect on domestic gas supplies leading to much lower gas prices
- 2011: LNG exports begin to get serious attention in the U.S. and Canada

Figure 20. Major sources of incremental natural gas supply, 2002-2025 (trillion cubic feet)



EIA AEO 2003

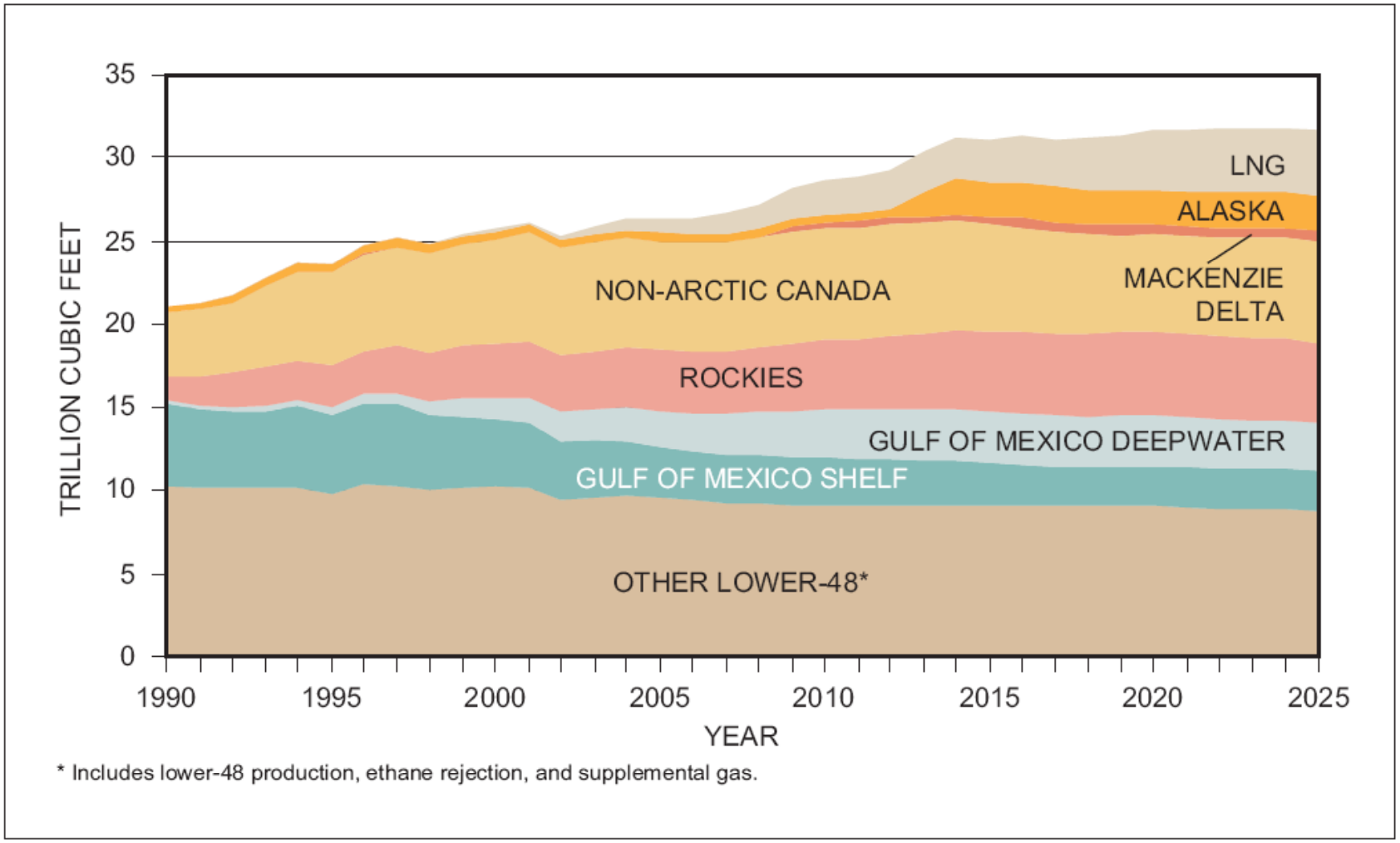
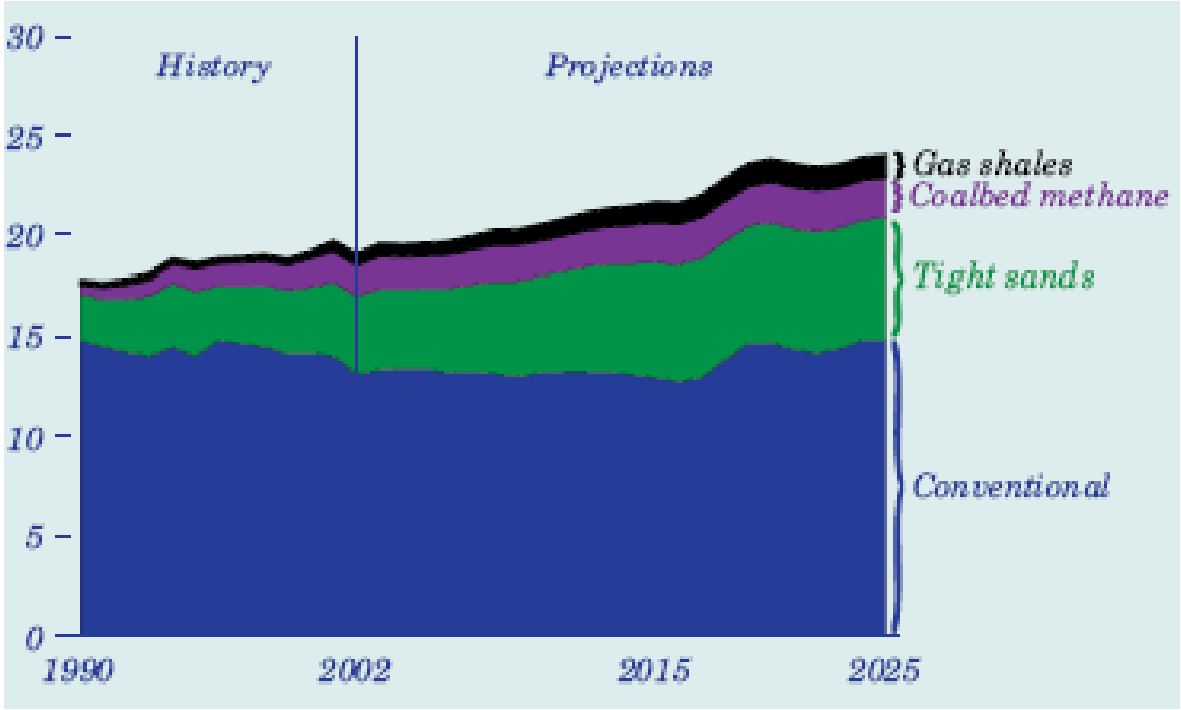


Figure 38. U.S. and Canadian Natural Gas Supply

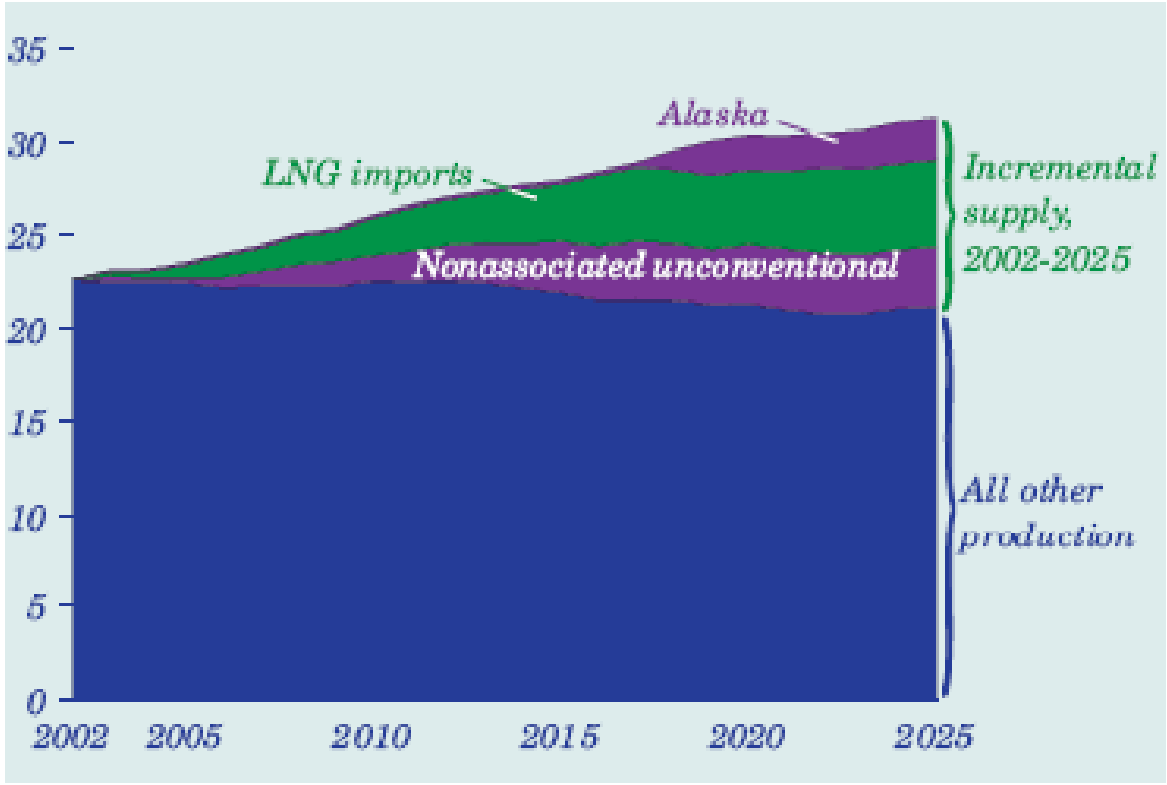
NPCC 2003

Figure 14. Lower 48 natural gas production by resource type, 1990-2025 (trillion cubic feet)



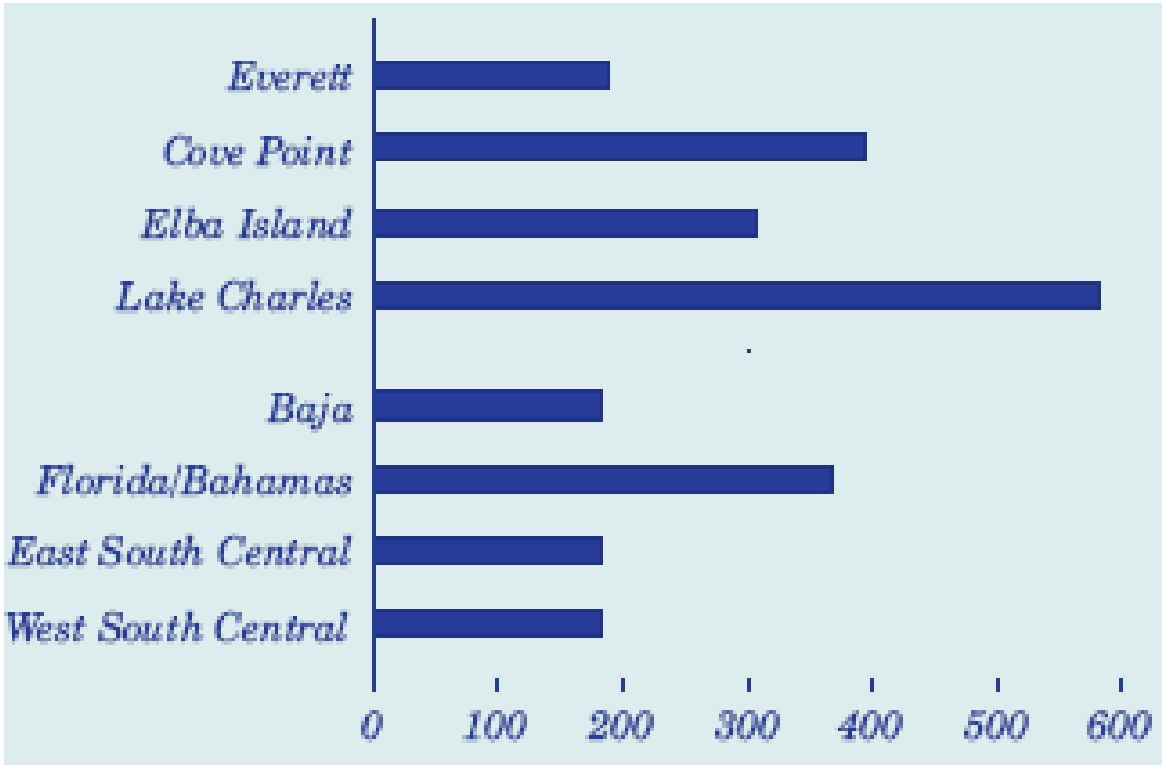
EIA AEO 2004

Figure 17. Major sources of incremental natural gas supply, 2002-2025 (trillion cubic feet)



EIA AEO 2004

Figure 21. Projected LNG imports by terminal and region in the reference case, 2025 (billion cubic feet)



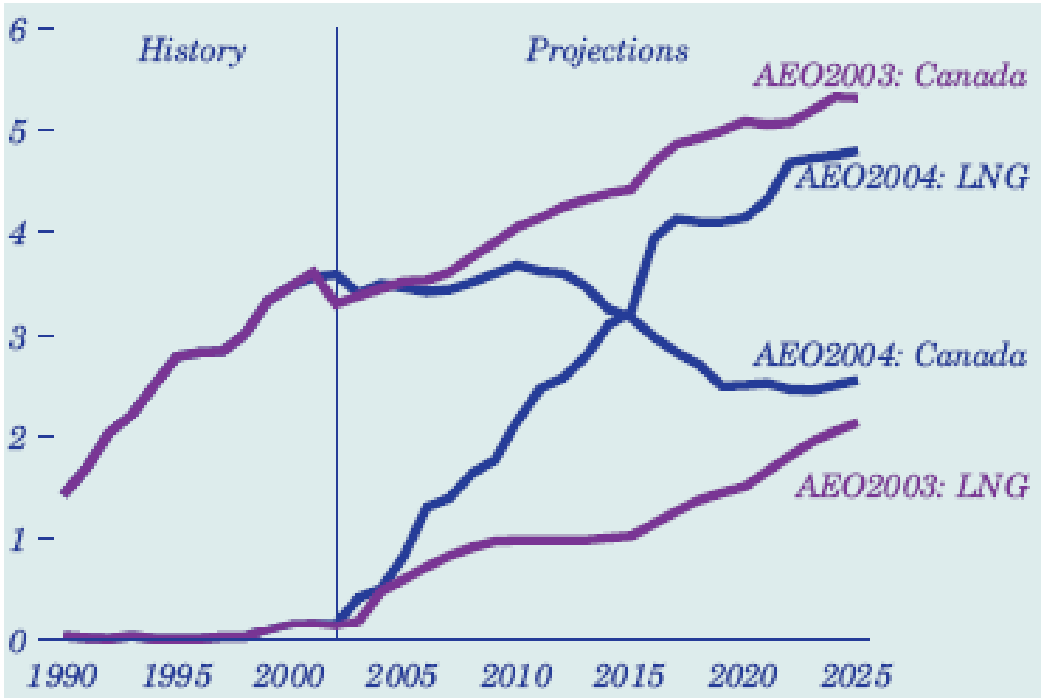
EIA AEO 2004

Table 12. North American LNG regasification proposals as of December 1, 2003 (million cubic feet per day)

Project	Owners	Location	Start year	Capacity added
West Coast				
Terminal GNL Mar Adentro de B.C.	ChevronTexaco	Baja California, Mexico (offshore)	2007	750
Tijuana Regional Energy Center	Marathon/Golar LNG/Grupo GGS	Baja California, Mexico	2006	750
Sound Energy Solutions	Mitsubishi	Long Beach, California	2007	700
Terminal LNG de Baja California	Shell	Baja California, Mexico	2007	1,000
Energia Costa Azul LNG	Sempra Energy	Baja California, Mexico	2007	1,000
Crystal	Crystal Energy	Oxnard, California (offshore)	2006	600
Tractebel Mexico	Tractebel	Lazaro Cardenas, Mexico	2007	500
Cabrillo Port LNG	BHP Billiton	Oxnard, California (offshore)	2008	1,500
Florida/Bahamas				
Ocean Express LNG	AES	Ocean Cay, Bahamas	2006	850
Freeport	El Paso	Freeport Grand Island, Bahamas	2007	500
Calypso	Tractebel Bahamas LNG	Freeport Grand Cayman, Bahamas	2007	832
Gulf Coast				
ExxonMobil LNG	ExxonMobil	Quintana Island, Texas	2007	1,000
Sabine Pass/Cheniere	Cheniere	Sabine Pass, Texas	2008	2,000
Port Pelican	ChevronTexaco	Louisiana (offshore)	2007	1,600
Cameron LNG	Sempra Energy	Hackberry, Louisiana	2007	1,500
Altamira	Shell	Altamira, Mexico	2004	500
Corpus Christi LNG	Cheniere Energy	Corpus Christi, Texas	2008	2,000
ExxonMobil/Sabine Pass LNG	ExxonMobil	Sabine Pass, Texas	2008	1,000
Liberty	HNG Storage/Conversion Gas	Cameron, Louisiana	2007	3,000
Main Pass Energy Hub	Freeport-McMoRan Sulphur	Gulf of Mexico (offshore)	2006	1,500
Gulf Landing	Shell	West Cameron, Louisiana (offshore)	2008-2009	1,000
Vermilion 179	Conversion Gas Imports	Louisiana	2008	1,000
Mobile Bay LNG	ExxonMobil	Mobile Bay, Alabama	2008	1,000
Freeport LNG	Freeport, Cheniere, Contango	Freeport, Texas	2006	1,500
Energy Bridge	El Paso	Floating Dock (offshore)	2005	500
East Coast				
Canaport	Irving Oil/Chevron Texaco	Canaport, New Brunswick, Canada	2006	500
Weaver's Cove	Poten	Fall River, Massachusetts	2007	400
Access Northeast Energy	Access Northeast Energy	Bearhead, Nova Scotia, Canada	2008	500
Fairwinds LNG	TransCanada, ConocoPhillips	Harpswell, Maine	2009	500
Providence LNG	Keyspan, BG LNG Services	Providence, Rhode Island	2005	500
Crown Landing	BP	Logan Township, New Jersey	2008	1,200
Somerset LNG	Somerset LNG	Somerset, Massachusetts	2007	430

EIA AEO 2004

Figure 20. U.S. net imports of LNG and Canadian natural gas, 1990-2025 (trillion cubic feet)

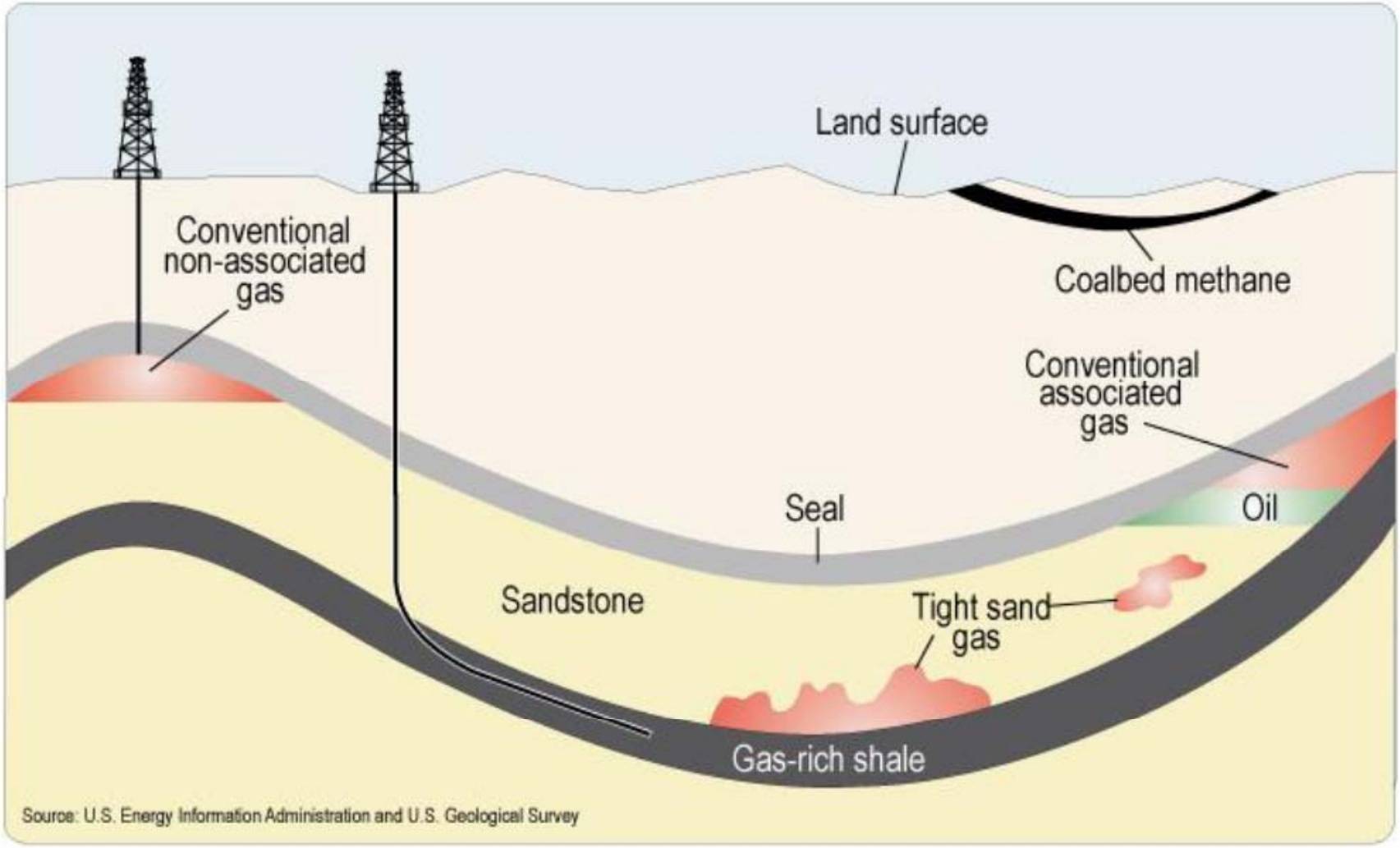


EIA AEO 2004

North American shale plays

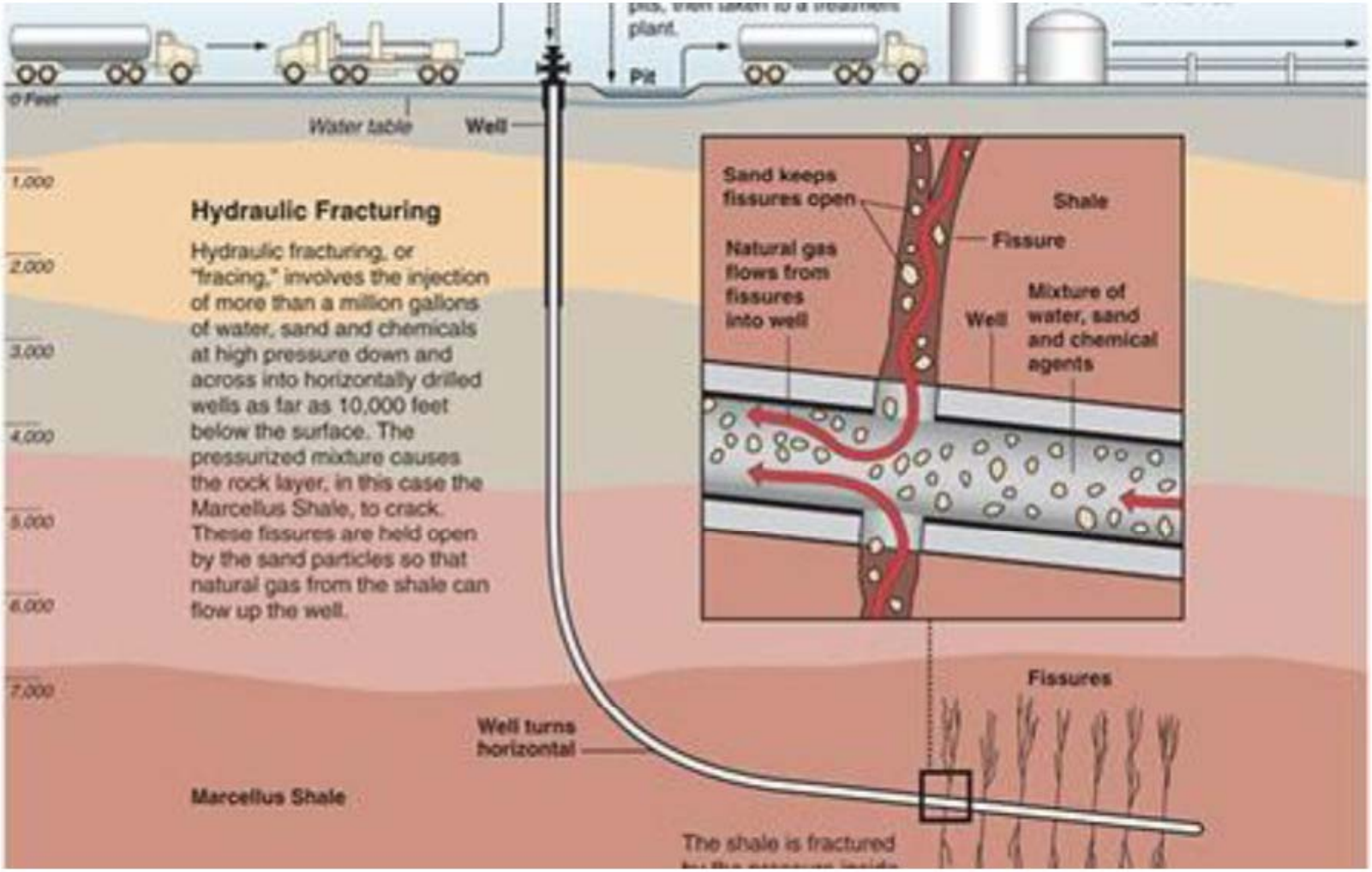
Source: U.S. Energy Information Administration based on data from various published studies. Canada and Mexico plays from ARI.

Underground sources of natural gas



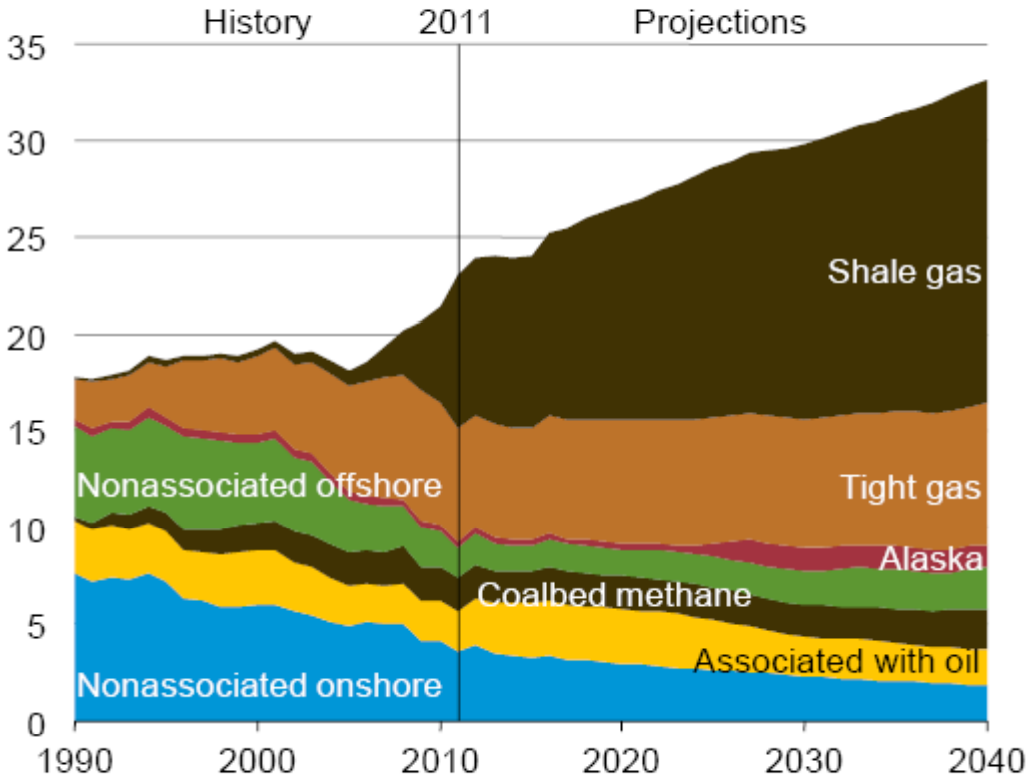
Source: modified from U.S. Geological Survey Fact Sheet 0113-01

Diagram of a typical hydraulic fracturing operation



Source: ProPublica, http://www.propublica.org/special/hydraulic-fracturing-national

Figure 3. U.S. dry natural gas production by source, 1990-2040 (trillion cubic feet)



EIA AEO 2013 Early Release (December 2012)

Figure 12. Electricity generation by fuel, 1990-2040 (trillion kilowatthours per year)

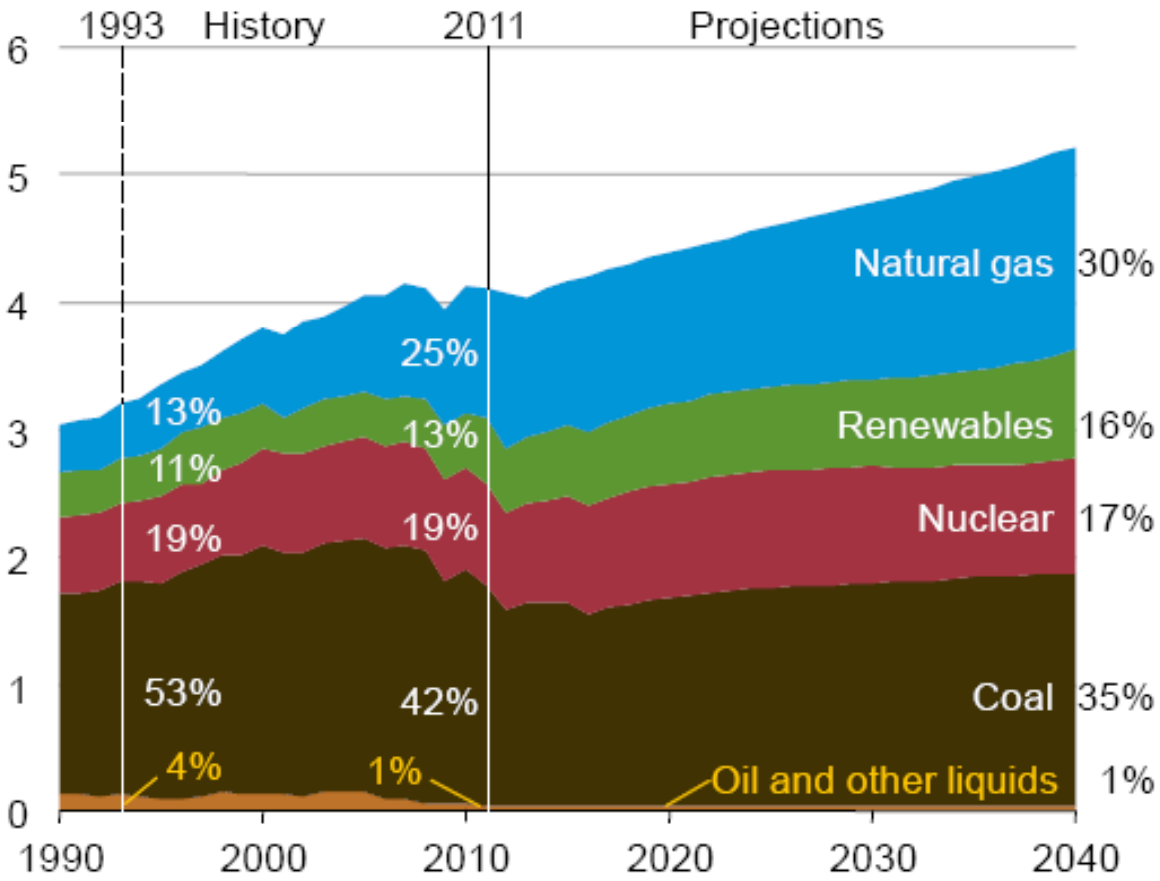
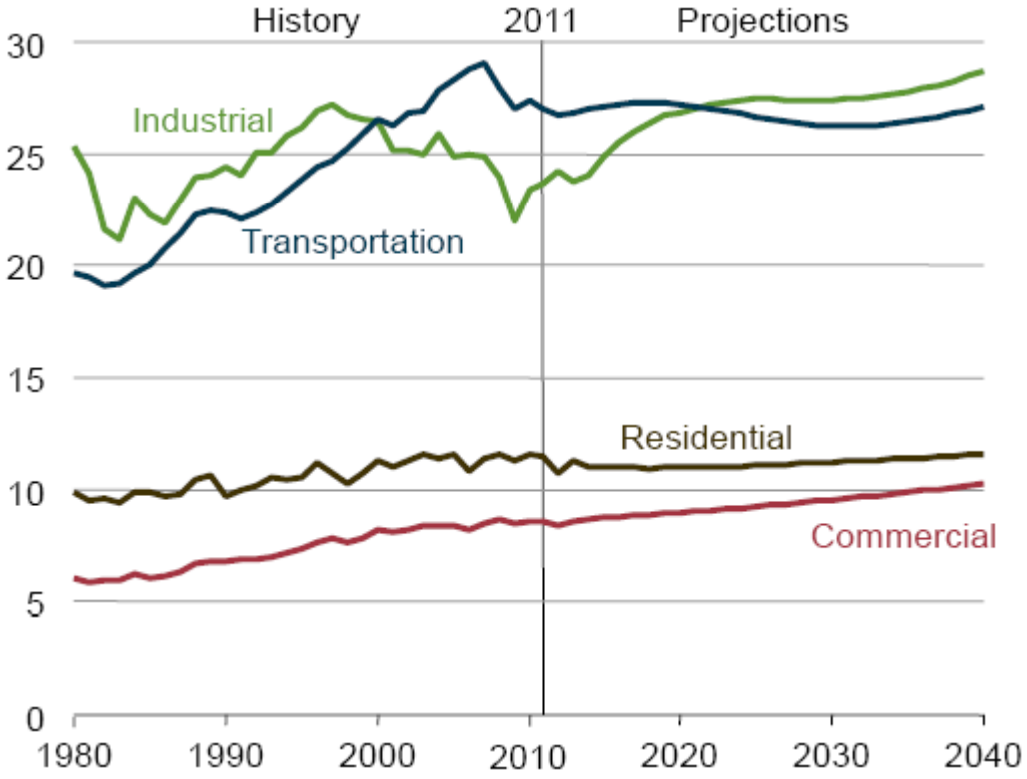
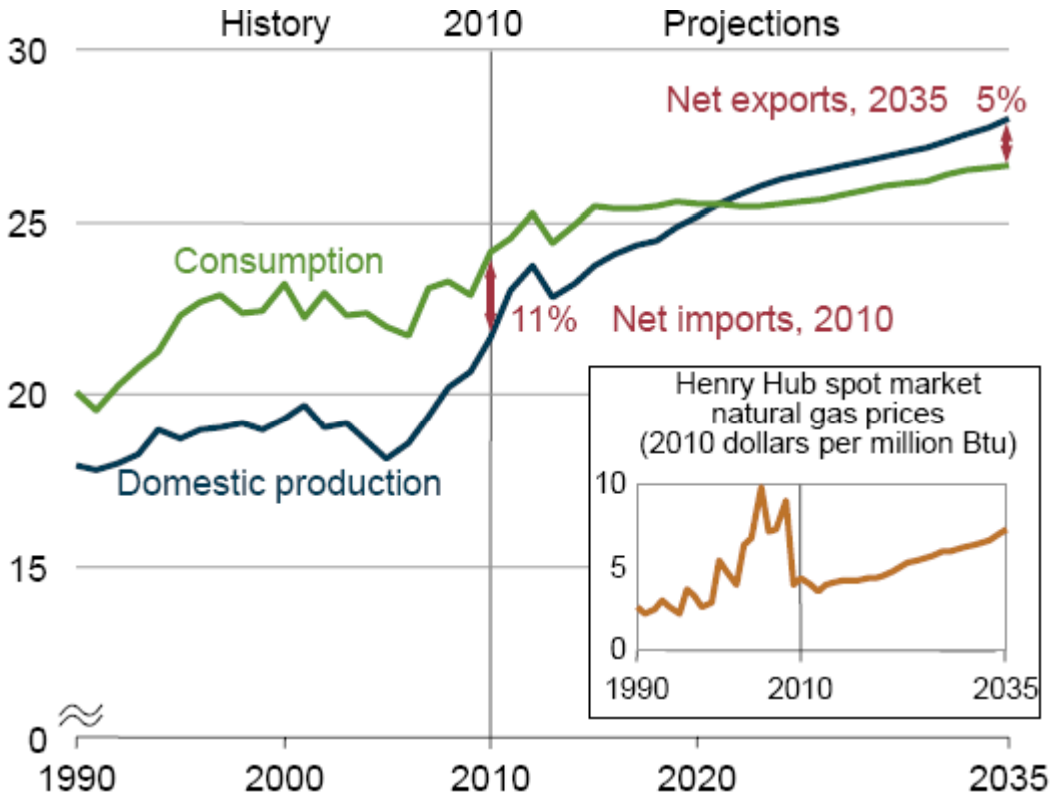


Figure 6. Delivered energy consumption by sector, 1980-2040 (quadrillion Btu)



AEO 2013 Early Release

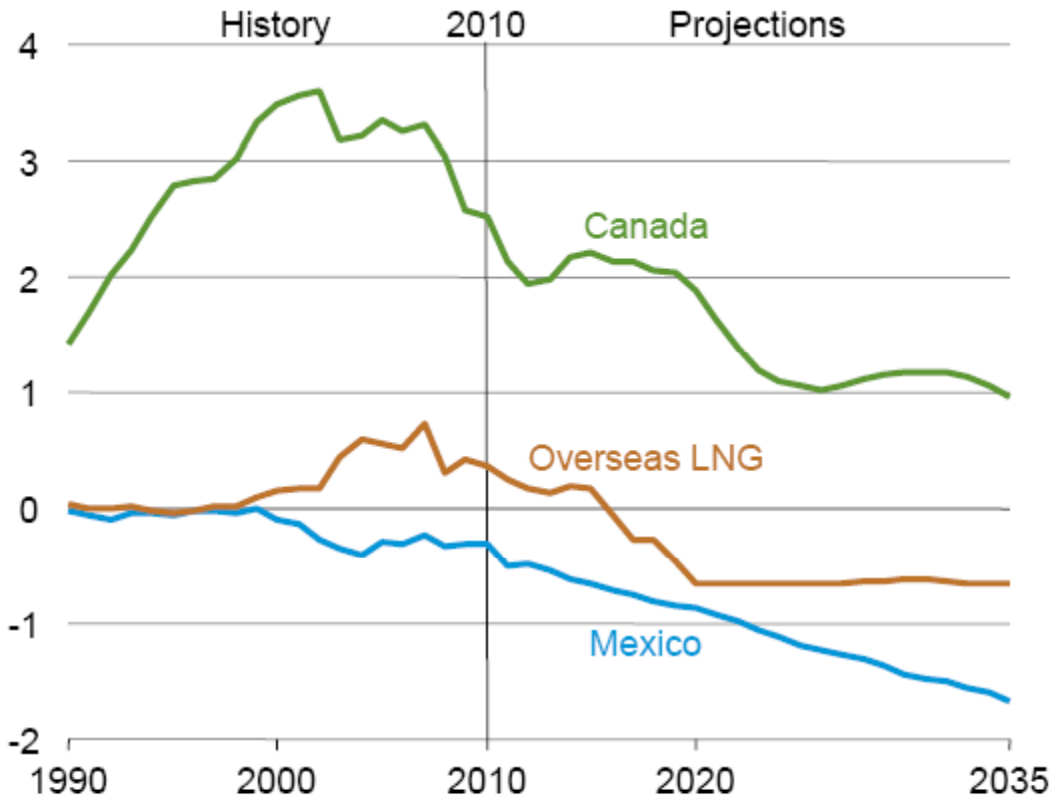
Figure 4. Total U.S. natural gas production, consumption, and net imports, 1990-2035 (trillion cubic feet)



Source: EIA AEO 2012

The U.S. becomes a net natural gas exporter

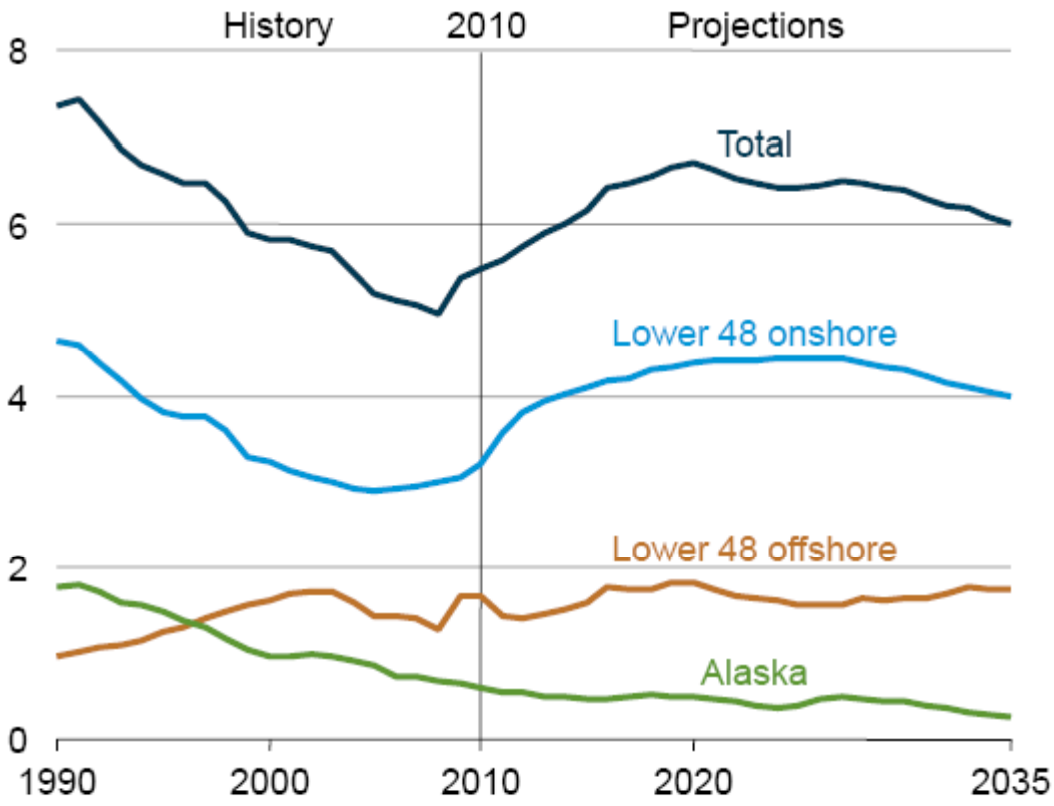
Figure 109. U.S. net imports of natural gas by source, 1990-2035 (trillion cubic feet)



Source: EIA AEO 2012

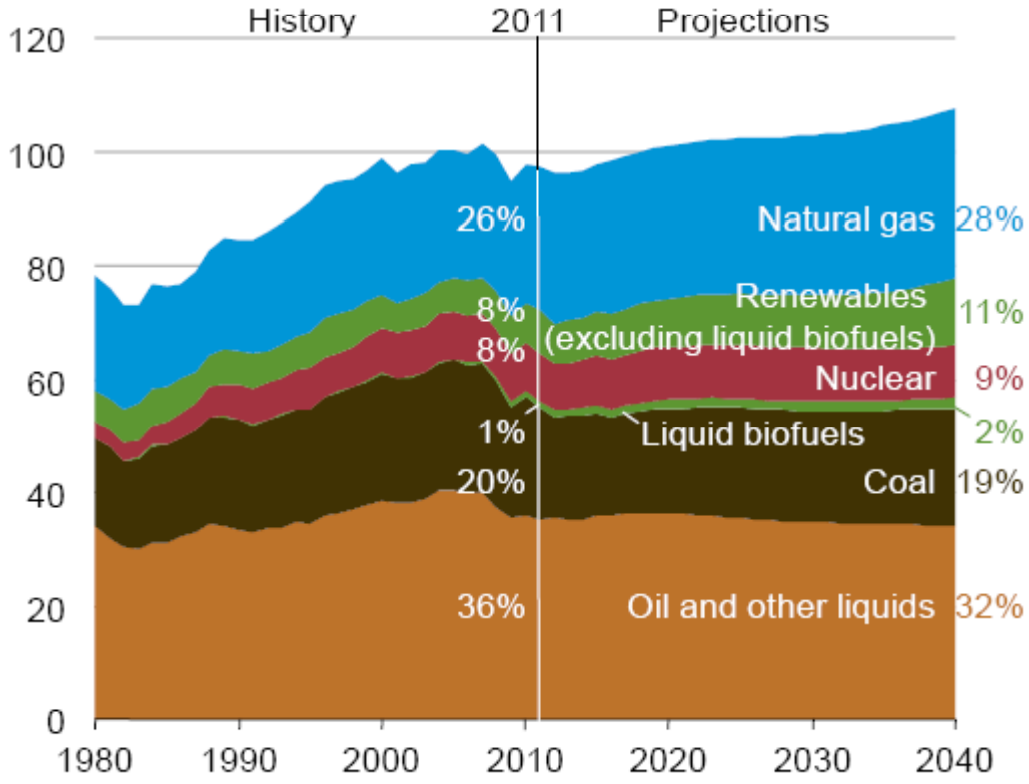
U.S. crude oil production increases, led by lower 48 onshore production

Figure 112. Domestic crude oil production by source, 1990-2035 (million barrels per day)



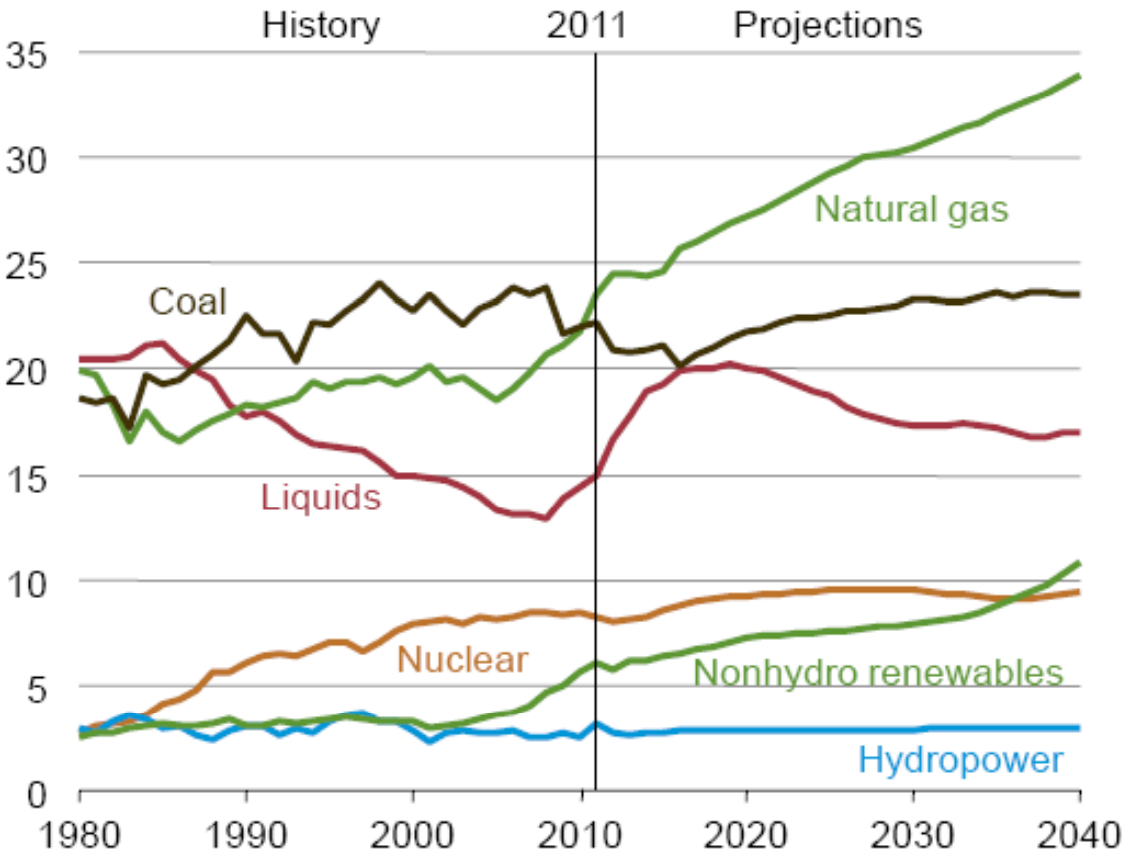
Source: EIA AEO 2012

Figure 7. U.S. primary energy consumption by fuel, 1980-2040 (quadrillion Btu per year)



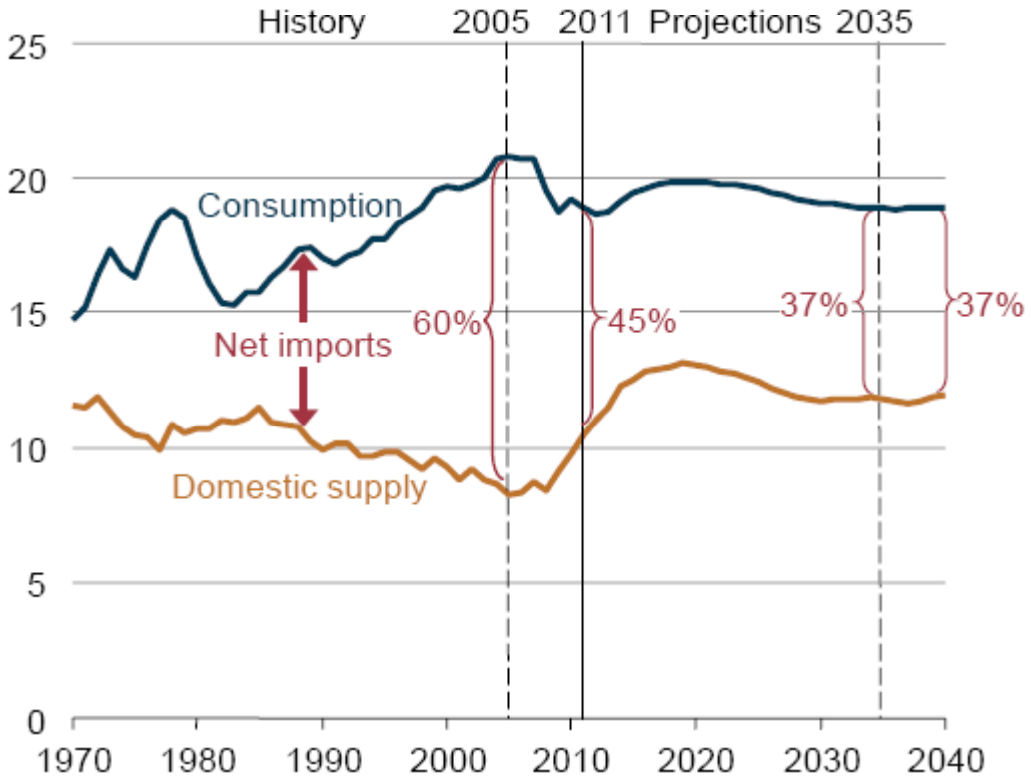
2013 AEO Early Release (December 2012)

Figure 10. U.S. energy production by fuel, 1980-2040 (quadrillion Btu)



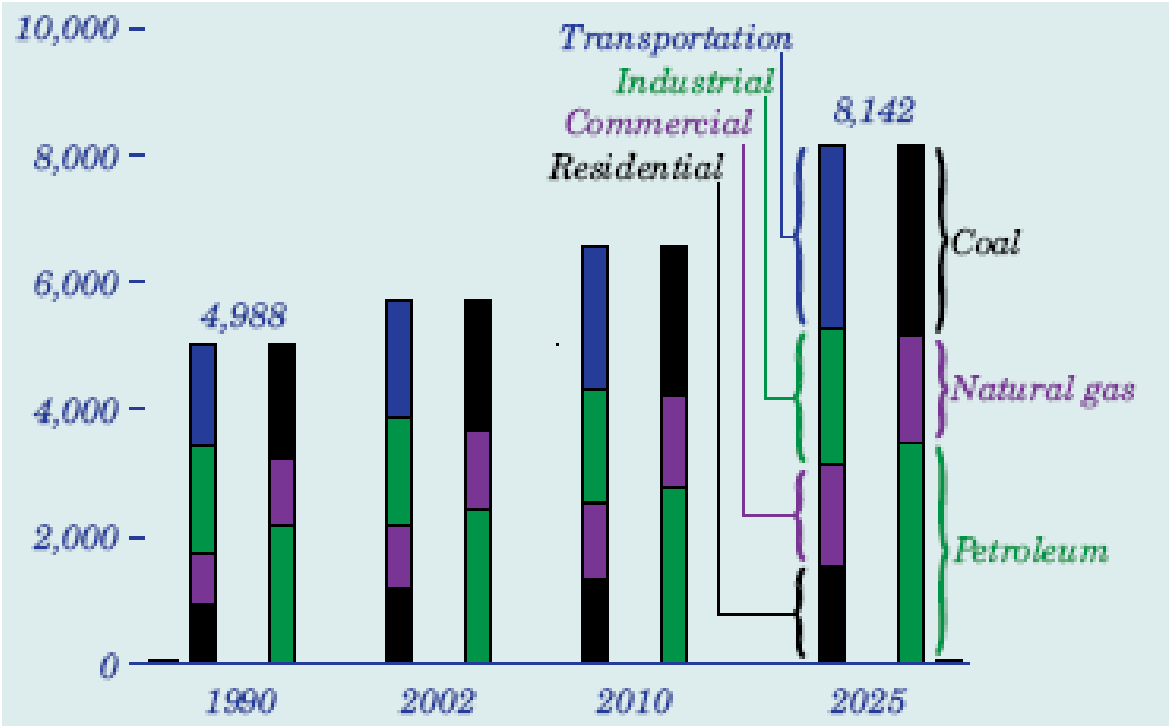
2013 AEO Early Release

Figure 11. U.S. liquid fuels supply, 1970-2040 (million barrels per day)



2013 AEO Early Release (December 2013)

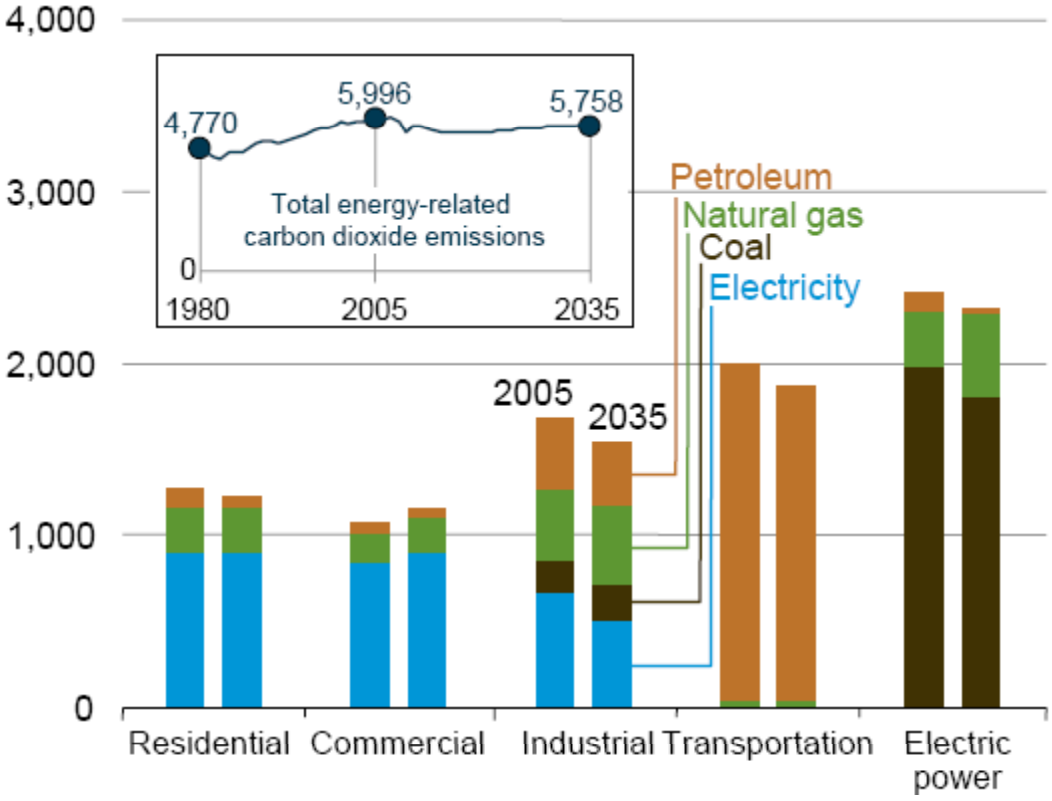
Figure 115. Carbon dioxide emissions by sector and fuel, 1990-2025 (million metric tons)



EIA AEO 2004

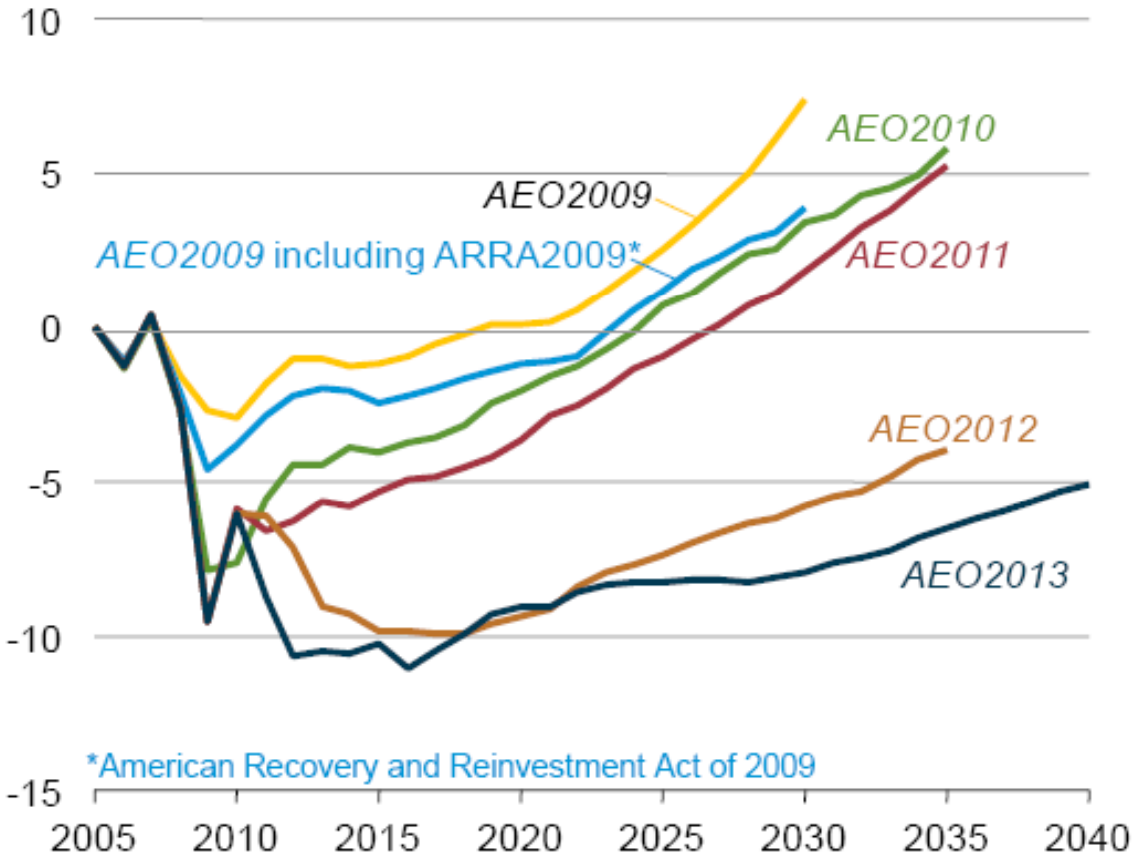
Projected energy-related carbon dioxide emissions remain below their 2005 level

Figure 122. U.S. energy-related carbon dioxide emissions by sector and fuel, 2005 and 2035 (million metric tons)



Source: EIA AEO 2012

Figure 13. U.S. energy-related carbon dioxide emissions in recent *AEO* Reference cases (percent change from 2005)



2013 AEO Early Release (December 2013)

Figure 10: 2010 LNG Trade (Tcf)

From\To	Africa	Canada	China/ India	C&S America	Europe	FSU	Korea/ Japan	Middle East	Oceania	Sakhalin	Southeast Asia	U.S.	**Total Exports**
Africa		0.03	0.05	0.31	1.33		0.24	0.21			0.07	0.31	**2.54**
Canada													**0.00**
China/India													**0.00**
C&S America		0.00		0.01	0.02		0.00					0.01	**0.05**
Europe				0.01	0.11		0.05	0.01			0.00		**0.18**
FSU													**0.00**
Korea/Japan													**0.00**
Middle East		0.01	0.44	0.08	1.15		1.28	0.10			0.15	0.08	**3.29**
Oceania			0.17				0.62				0.04		**0.83**
Sakhalin			0.02				0.39	0.00			0.02		**0.43**
Southeast Asia			0.14	0.06			1.92	0.01			0.21		**2.34**
U.S.							0.03						**0.03**
Total Imports	**0.00**	**0.04**	**0.81**	**0.47**	**2.61**	**0.00**	**4.53**	**0.34**	**0.00**	**0.00**	**0.49**	**0.40**	**9.70**

Source: "The LNG Industry 2010," GIIGNL.

Global Production ~ 120 Tcf

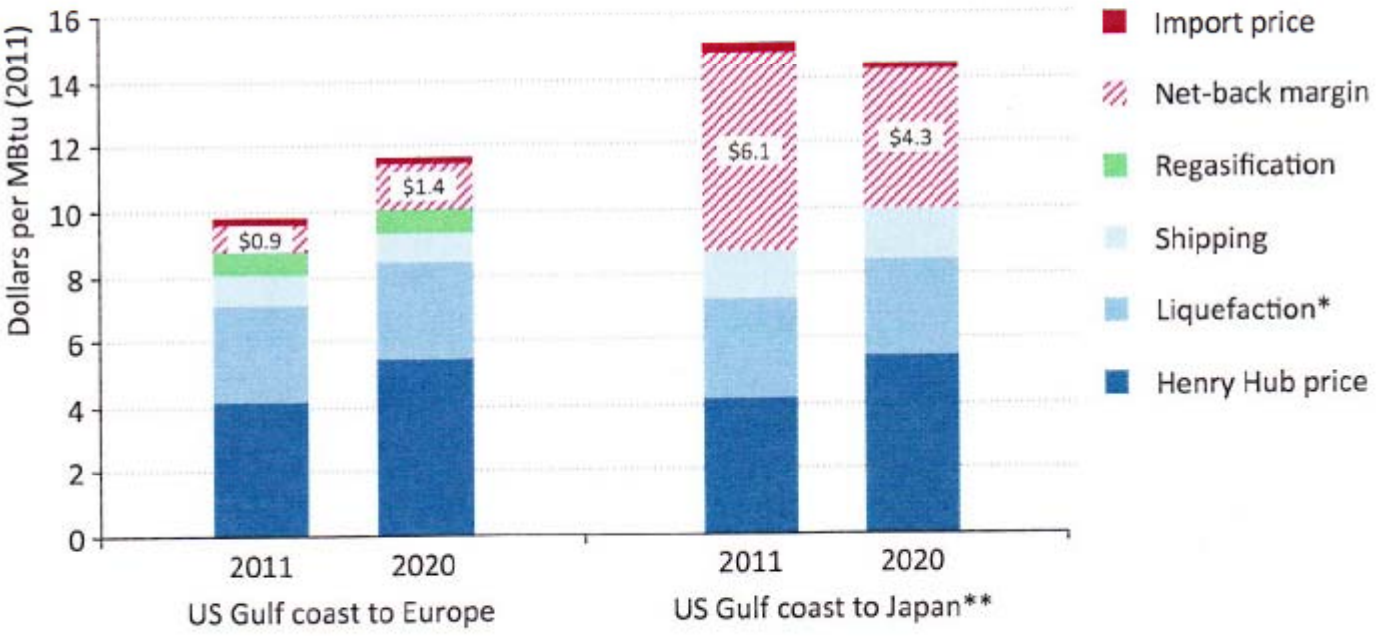
NERA Economic Consulting (2012)

Figure 16: Projected Wellhead Prices (2010$/MMBtu)

	2010	2015	2020	2025	2030	2035
Africa	$1.75	$1.89	$2.09	$2.31	$2.55	$2.81
Canada	$3.39	$3.72	$4.25	$5.20	$5.64	$6.68
China/India	$12.29	$12.86	$13.00	$13.25	$13.57	$13.51
C&S America	$2.00	$2.16	$2.39	$2.64	$2.91	$3.22
Europe	$9.04	$9.97	$10.80	$11.95	$12.39	$13.23
FSU	$4.25	$4.60	$5.08	$5.61	$6.19	$6.84
Korea/Japan	$14.59	$15.30	$15.47	$15.79	$16.19	$16.11
Middle East	$1.25	$1.35	$1.49	$1.65	$1.82	$2.01
Oceania	$1.75	$1.89	$2.09	$2.31	$2.55	$2.81
Sakhalin	$1.25	$1.35	$1.49	$1.65	$1.82	$2.01
Southeast Asia	$2.00	$2.16	$2.39	$2.64	$2.91	$3.22
U.S.	$3.72	$3.83	$4.28	$5.10	$5.48	$6.36

NERA Economic Consulting (2012)

Figure 4.3 ▷ Indicative economics of LNG exports from the United States

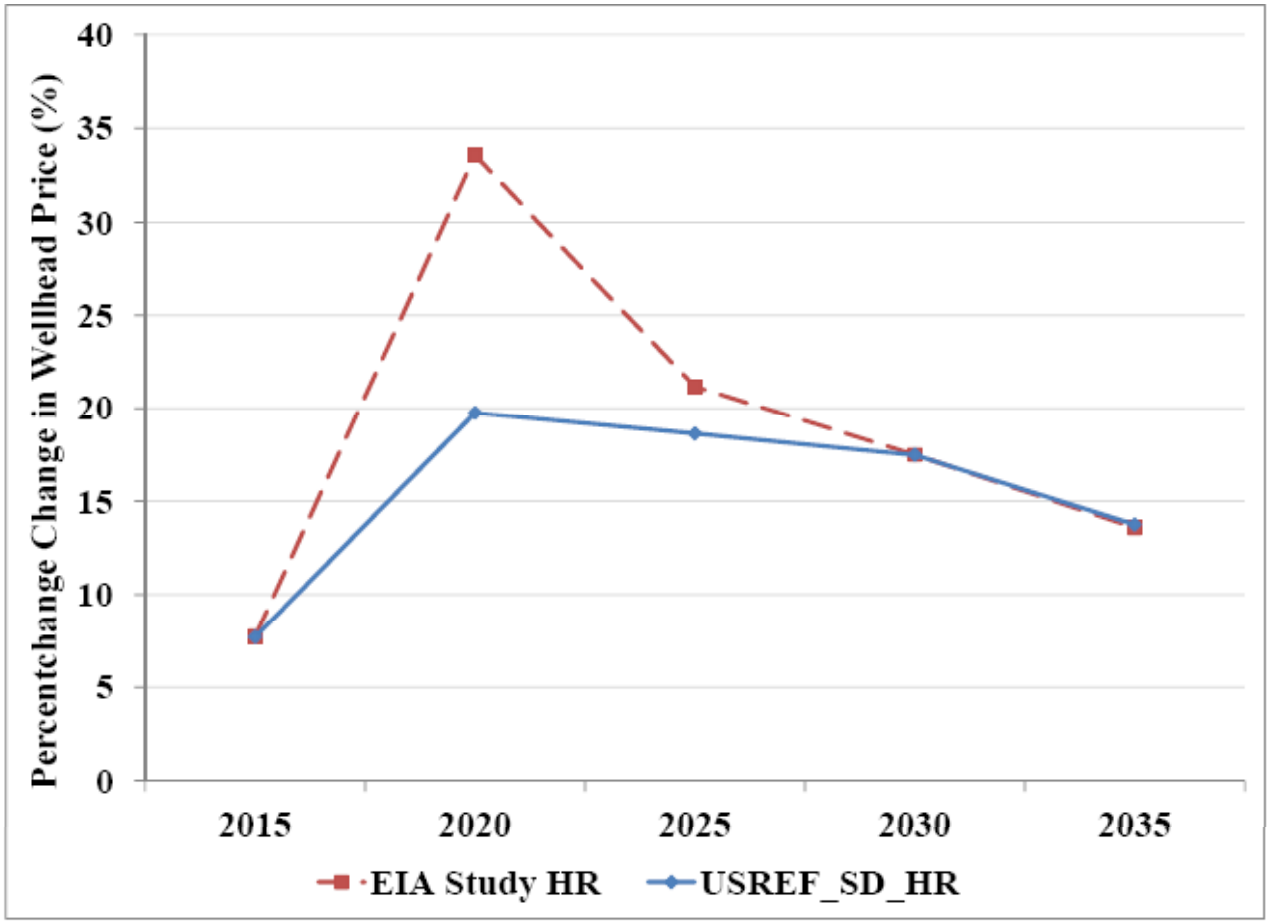


* Includes cost of pipeline transport to export terminal. ** Widening of the Panama Canal, due to be completed in 2014, will allow for more LNG tanker traffic.

Notes: LNG costs are levelised assuming asset life of 30 years and a 10% discount rate. The Japanese import price is for liquefied gas, so it does not include regasification.

Source: IEA World Energy Outlook 2012, p. 130

Figure 7: Comparison of EIA and NERA Maximum Wellhead Price Increases



NERA Economic Consulting (2013)

Environmental Issues

- Does “fracking” lead to contamination of drinking water aquifers with methane?
  - Hydraulic fracturing has been used in the oil and gas industry since at least 1947 and probably earlier
  - Fracking in shale deposits per se is unlikely to lead directly to drinking water contamination because it takes place thousands of feet below water aquifers
  - If there is contamination of drinking water it is much more likely due to failures in the vertical well casing, stray gas releases from conventional pools during vertical drilling, or surface spills of fracking fluids and drilling mud
  - The connection between shale gas production and groundwater contamination is weak and the studies that claim to document a connection are deficient --- little if any “before and after” analysis

Environmental Issues

- There are more important pathways leading to potential environmental impacts that are probably of more concern
  - Vertical well failures
  - Above ground spills of toxic fracking fluids and drilling mud and inadequate long-term disposal methods
  - Air emissions from drilling and support equipment
  - Increased truck traffic delivering equipment and disposing of wastes
  - Fresh water use
  - Earth tremors from below-ground disposal of fluids
  - Methane emissions
- Regulatory frameworks and capabilities vary widely from state to state and between states and federal government
- Best practices for regulatory frameworks have not been developed
- Lack of transparency by the industry has created suspicions that environmental impacts are not being taken into account adequately

Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use[1]

Katrina Jessoe[2] and David Rapson[3]

February 19, 2013

Abstract

Using data from a field experiment, we document that simple information feedback dramatically increases the price elasticity of demand for electricity, a setting where signals about quantity are coarse and infrequent. Households exposed only to temporary price increases reduce demand by 0 to 7 percent; adding information feedback induces reductions of 8 to 22 percent. This three standard deviation differential is significant and robust, and not attributable to awareness of price changes. Conservation extends beyond pricing events, providing evidence of habit formation, spillovers, and greenhouse gas abatement. Survey evidence points to learning as a likely mechanism for the treatment differential.

JEL: C93, D83, L94, Q41

Keywords: Information; Price Elasticity; Randomized Controlled Trial; Energy Demand

[1]We would like to thank Hunt Allcott, Jim Bushnell, Colin Cameron, Scott Carrell, Michael Carter, Meredith Fowlie, Koichiro Ito, Stephen Holland, Hilary Hoynes, Michael Price, Nancy Rose and Burkhard Schipper. This paper benefited from helpful comments by seminar participants at the CU Boulder, NBER, POWER, UC Davis, UC Energy Institute, and UCE3. Thanks to United Illuminating personnel for their important role in implementing this project. Tom Blake and Suzanne Plant provided excellent research assistance. Research support for this project was provided by UCE3 and United Illuminating. Rapson thanks the Energy Institute at Haas for support under a research contract from the California Energy Commission. All errors are our own.

[2]Department of Agricultural Economics, University of California, Davis, One Shields Ave, Davis, CA 95616; Phone: (530) 752-6977; Email: kkjessoe@ucdavis.edu

[3]Department of Economics, University of California, Davis, One Shields Ave, Davis, CA 95616; Phone: (530) 752-5368; Email: dsrapson@ucdavis.edu

1 Introduction

Traditional economic models assume that agents know the full choice set available to them, can costlessly compute strategy-specific payoffs, and have preferences that allow them to maximize their net economic benefit. Starting with Simon (1955) studies began to question these assumptions, recognizing that information may be costly to acquire and that there are limitations to human cognition and attention. Recent empirical studies have supported this view, showing that these features affect outcomes in market settings.[1] Much of the existing literature has focused on understanding what and how information is presented, emphasizing the role of price. However, in many settings there may also be a noisy signal about quantity, and less is known about the implications of this form of uncertainty on the efficiency of consumer decisions.[2] We create an experimental setting in which granular and accurate information on quantity is exogenously provided to some households exposed to price variation, allowing us to evaluate the effect of information on price elasticity. Providing simple real-time information to residential electricity users on their consumption increases the price elasticity of demand by three standard deviations.

The residential electricity sector is an ideal market to study this question for three reasons. First, this is an enormous industry known for its regulatory difficulties (e.g. market power) and external costs (e.g. pollution).[3] Second, the quantity of electricity consumption is shrouded from residential customers, in part because of coarse and infrequent billing.[4] Since electricity is inherently an input

[1]Consumers treat taxes featured and not featured in the posted price differently, becoming more price elastic when taxes are salient (Chetty et al. 2009). They are inattentive to opaque "add-on" costs such as shipping and handling fees (Hossain and Morgan 2006) and stock market earnings announcements, with inattention to financial news depending on the timing of reports (DellaVigna and Pollet 2008) and the number of competing stimuli (Hirshleifer et al. 2009). See DellaVigna (2008) for a comprehensive review of field evidence from this literature.

[2]Empirical work related to the role of information about quantity shows that consumers exhibit left-digit bias when considering mileage of used cars in their purchase decision (Lacetera et al 2012) and do not fully incorporate available information on cell phone usage when making choices about plans and future usage (Grubb and Osborne 2012).

[3]There are $370 billion per year in retail sales. This amounts to roughly two-thirds of US public expenditure on primary and secondary education in 2009. (U.S. Department of Energy 2010; National Center for Education Statistics 2011).

[4]Prices may also be confusing. Complex rate structures make it difficult for customers to know the marginal price. Recent work suggests that when faced with a non-linear price schedule, consumers often rely on a heuristic, responding more to average prices (Ito 2011). More frequent information on electricity prices

to household production rather than a final consumable, it lives in the background of our lives and makes the relevant unit of measure (kilowatt-hour) meaningless to most. Third, electricity customers traditionally exhibit low response to prices for reasons that are difficult to disentangle (Reiss and White 2005, Allcott 2011a, Ito 2011). It may be that, all else equal, low-elasticity households are unresponsive to prices because they are fully informed about how much it costs to use household appliances, equate marginal utility to marginal cost, and are rationally price inelastic. One alternative explanation, which we explore, is that the low elasticity is partly explained by the lack of full information about either price or quantity. If this is the case, the provision of information about price and quantity may lead consumers to make more efficient choices and in doing so change the way economists and policymakers think about price as a policy tool in the electricity market.

The ideal experiment to identify the incremental effect of information feedback on consumer price elasticity has not previously been implemented. This article describes results from such an experiment. We design and implement a randomized controlled trial (RCT) in which treatment households are exposed to exogenous price changes and (randomized) real-time feedback on both the price and quantity of electricity. To isolate the effect of price changes on demand, all treatment households were exposed to two- or four-hour long pricing events during which the price of electricity increased by 200 to 600 percent.[5] Households were informed of upcoming price changes via email, voicemail and in some cases text message. Some treatment households were (randomly) given in-home displays (IHDs) that provide real-time information on electricity prices and aggregate quantity consumed.[6] The in-home technology enables customers to inform themselves about electricity usage and prices at almost zero marginal cost. IHD installations occurred months before the first pricing event, and interested customers were free to use the devices to learn about how electricity actions translate into usage. Households assigned to a control group did not experience the temporary price changes, nor did they receive any technology.

---

has been shown to increase the short-run price elasticity (Allcott 2011a).

[5]This price change is a variant of what the industry refers to as "dynamic pricing". The increment is consistent with the magnitude of peak fluctuations in the wholesale electricity market.

[6]In-home displays are part of The Home Area Network (HAN). This is part of the household-facing layer of technology that transmits real-time information on electricity usage and prices from the meter to both the utility and household, offering consumers the potential to access more information and exert more control over their energy use.

Results from the experiment support the hypothesis that rich information about quantity plays an important role in consumer decisions. Conditional on changes in price, households exposed to information feedback consistently exhibit price elasticities that are roughly three standard deviations larger than those without feedback. Households in the price-only group reduce their usage by between 0 and 7 percent on average during pricing events, relative to control. In contrast, those exposed to the same price changes but who also have IHDs, exhibit much larger usage reductions of 8 to 22 percent. It is tempting to compare these estimates to the 1-to-3 percent overall usage reductions that have been achieved in other settings.[7] However, the periods over which our large treatment effects are observed are much shorter in duration (2 to 4 hours versus monthly usage), so the aggregate effect is likely comparable (though difficult to measure precisely in the context of the present research design). The benefit of highly targeted price changes is to improve grid efficiency for short periods of time, and our main result shows that the incremental contribution of real-time information is remarkable in achieving this.

We also investigate the role of heterogeneity within our sample using data collected from household surveys. A common hypothesis is that households are inattentive to electricity prices because they are not aware of them, perhaps because price changes are not salient. We attempt to control for this in the experimental design, and also test ex post if the information treatment differential can be explained by an increase in the awareness of pricing events for households with IHDs. While awareness of price events is correlated with overall response, it does not explain why informed households are more price elastic than their uninformed counterparts. Instead, our empirical evidence suggests that experience with IHDs facilitates consumer learning, helping households to make better decisions when confronted with high prices. Consumers who look at the information device more frequently are significantly more responsive to price. This is consistent with a framework in which experience with the IHDs allows consumers to better understand how their day-to-day actions translate into electricity usage. It also fits with earlier empirical work showing that frequent market experience can reduce market anomalies (List 2003, Feng and Seasholes 2005).

The dynamic effects of treatment, both within an event day and on non-event days, cause our

[7]See for example Jacobsen et al. 2010, Allcott 2011b, and Harding and Rapson 2012.

estimates to be a lower bound of the primary treatment effects as well as of the information treatment differential. Within a day, the conservation effect observed during pricing events spills over into adjacent hours for households with information feedback. In the longer run, an evaluation of trends in usage over the days of the summer reveals that households in both the price and price+IHD group are forming conservation habits, even when events are not occurring. The longer run behavioral change adds another dimension to our understanding of consumer choice, providing further evidence of learning. The combined dynamic effects imply a social benefit in the form of greenhouse gas abatement.

The main result of this paper implies that incomplete information may be causing large efficiency losses, suggesting a potential role for government intervention to correct the market failure.[8] Further exacerbating inefficiency in this industry is the well-documented mismatch between wholesale and retail prices.[9] If consumers were to face a critical peak pricing (CPP) program similar to the one we implement, information feedback via an IHD would augment the efficiency gain. To assess the net benefit of a large-scale IHD deployment, we provide an estimate of the payback period from private and public investment in the feedback devices. In this context, an IHD pays for itself privately in 6-20 years on average. The social benefits of deploying IHDs to one-third of U.S. households yield an upper bound payback period of 7-9 years.

Our experimental design produces internally valid estimates of the treatment effect.[10] At the same time, the sample was drawn from a potentially-select group of households; each household that volunteered for our study is enrolled in paperless billing and also has a broadband Internet connection. The main results persist even after weighting the sample to reflect the national income distribution, demonstrating that selection on income is not the only factor. Nonetheless, our results must be interpreted in this context. There remains a strong need for our experimental design (or similar) to be implemented on a variety of sub-populations with different characteristics.

[8]Greenwald and Stiglitz (1986) demonstrate that economies characterized by incomplete information are not, in general, constrained Pareto efficient.

[9]Examples include Borenstein 2002, Borenstein 2005, and Borenstein and Holland 2005.

[10]To our knowledge, the only other peer-reviewed study that studies both information and price in electricity is Allcott (2011a). Information in that setting is in the form of a price orb that changes color as prices increase, but gives no feedback about quantity. Another contrast with this study is that the price orb program designed by Commonwealth Edison was not free from selection concerns.

The paper proceeds as follows: section 2 presents a simple conceptual framework to motivate the empirical approach; section 3 describes the experimental setting and research design, and section 4 lays out the data; the empirical methods and results are presented and discussed in section 5; section 6 describes dynamic effects and section 7 follows with a discussion of welfare and policy implications; section 8 concludes.

2 Conceptual Framework

Some theoretical features of our setting motivate the empirical analysis. Consider a household that derives utility from the use of energy services such as air conditioning or illumination. The quantity of energy inputs (kWh) required to produce these services is not well understood. While consumers receive monthly bills summarizing aggregate monthly electricity usage, these bills do not detail energy usage by service, nor do they display the energy-intensity of these services. As a result, it is difficult if not impossible for even the most attentive consumer to perform a marginal cost-benefit analysis comparing energy services. Further, the disconnect between perceived usage and services causes consumers to hold incorrect beliefs about how changes in the production of energy services impact energy usage.

Assume that consumers have full knowledge about the marginal benefit of household energy services but are uncertain about the marginal costs. This initial uncertainty may be about the price of electricity, the quantity of electricity required to produce a service, or both. Ex ante (before the introduction of feedback), consumers equate the expected marginal benefit of each activity to its expected marginal cost. The expected marginal cost of an electricity service will depend on a consumer's prior beliefs about how each potential service maps into electricity usage. Of course, the prior may be incorrect; some activities may be more energy-intensive than others, relative to expectations.

Once a consumer is exposed to feedback, information about the mapping between usage and services is revealed. She will begin to learn how choices about electricity actions translate into usage and

expenditure. Ex post, this information may reveal "mistakes" in optimization (even conditional on price), inducing adjustments. Over time, prior beliefs will be updated to incorporate information contained in the feedback, and in the limit the posterior distribution will converge to the true mapping.

As an example of the general case, consider a consumer whose electricity use is comprised of a fixed and variable component. The fixed portion of usage comprises baseline household services such as the continuous operation of the refrigerator and deep freezer, and some baseline level of temperature control and illumination. The discretionary component represents the margins of adjustment that are perceived to be available should a customer want to increase or decrease her electricity consumption.[11] Let X denote overall electricity use,

$$X = \bar{h} + h \tag{1}$$

where $\bar{h}$ and h describe the fixed and variable component respectively. Define the share of total usage that is discretionary as $\alpha \equiv \frac{h}{X}$, and assume that uncertainty only exists with respect to the value of α. For example, households may perceive standby power (also know as "phantom load") as fixed, whereas it is in fact variable.[12] This will cause them to under-estimate their true α.

We now posit certain features that likely describe the relationship between the price elasticity of demand for electricity, ε, and α. Let the demand elasticity be defined as a function $f(\alpha)$,

$$f(\alpha) = \varepsilon \equiv \frac{\partial log X}{\partial log p} \tag{2}$$

where p is the price of electricity. First, $f()$ will be monotonically increasing in the perceived level of α. As $\alpha \to 0$, demand becomes perfect inelastic since the perceived share of discretionary usage approaches zero. On the other extreme, as $\alpha \to 1$, all usage is perceived as discretionary, leaving all margins of adjustment available to the consumer when responding to price changes. As to the

[11] In reality, the distinction between discretionary and non-discretionary usage is blurred. At a high enough price, all consumption is discretionary.

[12] Standby power is the electricity usage attributable to appliances that are plugged in and not turned on, but are still drawing load. "Phantom load" is estimated to account for 10 percent of U.S. residential electricity use. See http://standby.lbl.gov/ for more details.

curvature of $f()$, it is intuitive that for "high enough" values of α, $f()$ will be locally concave. A higher fraction of discretionary use leads to more margins on which to respond to price changes, but the incremental contribution of these margins is decreasing.

To determine how behavior will change as information about the "true" value of α is revealed, one must consider prior beliefs about α. Denote by α^0 the true share of discretionary usage, and by $\tilde{\alpha}$ the consumer's uninformed belief. Depending on the initial relationship between the uninformed belief and the true share of discretionary usage, the effect of information on the price elasticity of demand is ambiguous. If $\tilde{\alpha} > \alpha^0$, then information that moves consumer beliefs closer to α^0 will decrease elasticity; if $\tilde{\alpha} < \alpha^0$, then information will cause the elasticity to increase. However, the magnitude of these effects is asymmetric. If the distribution of prior beliefs is symmetric around the truth (that is $E(\tilde{\alpha}) = \alpha^0$), then the average demand elasticity will increase in the presence of new information. Even this simplified and very stylized framework reveals the complexity of factors that may predict the role of information in price response. Thus, we believe an empirical inquiry is appropriate to answer the question: does information increase price elasticity?

3 Research Design

In partnership with a utility, we designed and implemented a *framed field experiment* (or randomized controlled trial) that introduced short-term price increases and real-time information to a sample of residential electricity customers in Connecticut.[13] The treatment events occurred in July and August of 2011. In this section, we describe the empirical setting, our sample, and the research design.

The United Illuminating Company (UI) is an investor-owned utility supplying 320,000 customers

[13]Home Area Network (HAN) devices such as the displays evaluated in this study, are being tested by utilities in a number of pilot settings, often in conjunction with dynamic pricing. These pilots have led to several industry research studies, some of which share features of ours. However, utility-run experiments are often constrained by the desire to avoid treating households differently, which can be perceived as unfair. From a research design perspective, this leads to selection bias since households are able to affect the treatment that they receive. Examples can be found in Faruqui and Sergici (2010), Faruqui et al. (2012), and Herter (2012).

in the New Haven and Bridgeport areas of Connecticut. Currently, households in UI's territory are on one of two pricing plans: a flat rate of \$0.21/kWh or a time-of-use (TOU) rate which charges a "peak" price of \$0.27/kWh on weekdays from noon until 8 pm, and an off-peak price of \$0.18/kWh at all other hours. UI has been utilizing Smart Grid technology for decades. In the late 1970s, UI replaced traditional analog meters with automated meter readers (AMR) for all of their commercial and industrial customers, enabling it to become one of the first utilities to offer TOU pricing.[14] In 2010 the Department of Public Utility and Control (DPUC) provided funding (i) for the installation of the most recent generation of smart meters called Advanced Metering Infrastructure (AMI) and (ii) to test the potential of smart technologies, including in-home displays and other HAN devices to shed load and conserve energy. It is within the context of this initiative that we designed the experiment.

To be eligible for participation in the pilot a customer needed to reside in a townhouse or single family home, have a broadband internet connection, and sign and return an end-user agreement indemnifying UI against litigation risk.[15] As a participation incentive, the utility offered households \$40: \$20 upon completion of a pre-survey prior to assignment to treatment and \$20 upon completion of a survey once the pilot ended. To recruit households into the HAN pilot, UI emailed 60,000 customers that had enrolled in paperless billing, indicating the likely presence of Internet in their home. It is well known that only a fraction of households receiving these sorts of marketing emails actually open them, let alone read them. We estimate that approximately 7,000 households opened the emails.[16] Of the 7,000 or so households that received the invitation to participate, we recruited 1,152 households (approximately 1 in 6) to participate in the project. A subset of these households, 437, form the sample for this study.[17]

[14] AMRs are an early generation of the smart meter that allows for one-way communication of information, from the home to the utility.

[15] Some aspects of device-to-utility communication were configured to occur wirelessly via the Internet. To maintain the integrity of our randomization, it was thus important that *all* of the participants have wireless. Any household that upon recruitment had a wireline broadband connection instead of a wireless connection was given a wireless router.

[16] We are unable to report the exact open rate (the fraction of delivered email messages that are actually opened) since return receipts were not collected. However, 12 percent is a-rule-of-thumb open rate used by industry experts. For example, see http://www.mailermailer.com/resources/metrics/2012/open-rates.rwp.

[17] This study represents a portion of a larger project that tests other aspects of the HAN.

3.1 Treatments

We randomly assigned households to one of three groups: control, price ("price-only", or occasionally "price"), and price-plus-information ("price+IHD"). Our decisions about cell size were guided by power calculations and a financial resource constraint.[18] A total of 207 households were assigned to the control group. These households (and all others in the pilot) received a mailing notifying them that they were in the pilot and informing them of their group assignment. They were also mailed an energy conservation pamphlet documenting "101 Ways to Conserve Electricity". The remaining 230 households were assigned to one of two treatment groups.

Price-Only Treatment: The 130 households in this group experienced 2 types of pricing events that varied in the magnitude of the price increase and the timing of event notification. The first event type, "DA", provided *day-ahead* notification that the per-kWh price of electricity would be increased by $0.50. These events mimicked a pricing policy that a utility might use to incentivize electricity conservation when high temperatures are expected in the following afternoon. The second type of event, "TM", sent notification *thirty minutes* before a $1.25 increase in the per-kWh price of electricity. A utility might implement this type of policy to reduce immediate risk in grid stability due to an unexpected decrease in generation (say, due to the failure of a baseload generating plant).[19] Ex-ante we cannot predict to which type of event households will be more responsive. While TM events send a much stronger price signal, households may not be able to respond to the price change within the short window of advance notification.

Between July 2011 and August 2011, three DA and three TM pricing events occurred. While all events occurred during peak hours, there was variation in the length and timing of events. Table (1) provides a description of each event including the start time, event duration, and the mean

[18] We directly measured the within- and across-household variance of electricity usage with the benefit of a secondary source of high-frequency usage data from existing UI customers, which is collected as part of UI's regular course of business. This enabled us to anticipate the statistical power of various candidate experimental designs.

[19] On July 22, 2011, a prolonged heat wave on the east coast was occurring and the peak wholesale price on the New England Independent System Operator's (NEISO) spot market climbed to nearly $0.60/kWh. If households on a flat rate faced a DA event on that day, the retail price would have been $0.71/kWh. A significant disruption of supply would have compounded strain on the grid, and prices may have easily approached levels on the order of magnitude of our experimental price changes.

and high temperature of the day. Households received notification of these pricing events by email, phone call and possibly text message, depending on their preference.[20]

Due to regulatory constraints, these price changes could not be reflected in the customer's monthly electricity bill. In addition, the utility wanted to ensure that no household was made worse off due to the price changes. To navigate these constraints while still achieving the intended marginal incentives, we provided customers assigned to the price treatment with an off-bill account initially credited with $100. The difference between the regulated price and the event price was multiplied by the quantity of electricity consumed during each event time period, and that amount was added to or subtracted from the household's off-bill account balance. It should be noted that the event price, not the regulated price, was conveyed to customers. Following an event, a household would receive an email, text or call reporting the balance in its off-bill account. By design, the marginal incentive facing consumers was equivalent to a change in their per kWh price. At the conclusion of the experiment, any balance remaining in the account was the customer's to keep.[21]

Price+IHD Treatment: The 100 households assigned to this treatment group experienced the pricing program described above, and also received real-time information about electricity prices, usage and expenditure. This information was provided via an in-home display (IHD), a portable device which can be mounted on a wall or placed on a counter, that displays in real-time electricity consumption, prices and bills. Figures (A-1) and (A-2) provide photos of the IHDs used in this study. These households also received computer software enabling them to log in to a web portal to monitor and view electricity usage, prices and expenditure.

The main difference between this treatment and the price treatment is that customers are able to view in real time the price of electricity, quantity of power being consumed, and their estimated monthly bill-to-date. The device removes the information acquisition costs involved in informing oneself how electricity-consuming actions into electricity usage and expenditure. The ability to

[20]Upon enrollment in the pilot, households were asked to choose how they would like to be informed of pricing events should they be assigned to this treatment group, and were able to select one or more of the alternatives.

[21]The unique implementation of price changes may raise concerns about construct validity. However, the central result of the paper relates to the differential response to prices between households with and without information feedback. Any concerns about construct validity would apply equally to these two groups.

view real-time usage provides customers with the opportunity to learn which appliances are heavy electricity users and which are not, thereby enabling them to more fully optimize in response to price changes.

Households received the IHDs and professional installation free of charge. The latter ensured that displays were set up correctly and activated, and maximized the likelihood that participants understood how to use them. Aside from the installation and set-up of the IHD, the vendor provided no services.

4 Data

In this section we discuss the effectiveness of the randomization, compliance and descriptive statistics. The primary data source is high-frequency meter data on household electricity usage. To collect these data, advanced meters were installed in all participating households.[22] During the experiment, electricity usage of all participants (including those that dropped out of the study, the so-called "non-compliers") was collected at 15-minute intervals. In addition to the usage data, the utility collected data confirming the receipt of event notification, as well as customer rate class (flat rate or TOU rate) and whether central air conditioning is present in their dwelling.

Households in the experiment completed two surveys: one prior to assignment to treatment (the "pre-survey") and another upon completion of the pilot (the "post-survey"). These surveys collected data on household demographic characteristics, housing unit characteristics, appliance ownership, and conservation-related actions. In the pre-survey, we also asked respondents about their propensity to be at home during the day. In the post-survey, we surveyed households on their awareness and understanding of pricing events, and the frequency with which they checked their IHDs. These questions allow us to investigate potential mechanisms and explanations for our primary estimates.

[22]Households that already had the older and less advanced Automated Meter Reading (AMR) devices did not receive the new AMI installations. Instead, their meters were adjusted to collect the same high-frequency data as the AMI meters.

4.1 Randomization and Compliance

Before exploring the impact of information on the price elasticity, we seek evidence to support the integrity of the randomization and that attrition is not imposing post-assignment selection bias. We first compare observable characteristics of treatment and control households. Table (2) presents descriptive statistics by treatment. Panel A includes all households who initially agreed to participate in the study and Panel B is restricted to households who completed the pilot ("compliers"). Thirty-eight households (approximately 9 percent) did not complete the pilot. These households either moved, requested to be removed from the study or failed to arrange an installation appointment for the IHD. Due to the difficulty of scheduling installation appointments for households assigned to the price+IHD group, attrition is not balanced across treatment groups. Of the 100 households assigned to the price+IHD treatment, 28 did not complete the study. In contrast, compliance was high in the non-technology groups (control and price-only). Only 4 of the 207 households assigned to the control group and 6 of the 130 households assigned to the price treatment moved or requested to be removed from the study. The concern raised by this asymmetry would be that success or failure to schedule an installation appointment is systematically related to the desire or ability to respond to treatment. This does not appear to be the case, and below we discuss how our empirical specifications account for this concern. First, though, we describe the sample.

As shown in Panel A of Table (2), peak hourly electricity usage averages at 1.42, 1.53 and 1.38 kWh in the control, price and price+IHD treatments, respectively. A comparison of mean usage between each treatment and control indicates that both peak and off-peak usage are statistically indistinguishable between groups if the sample includes households initially assigned to each treatment (see columns labeled "Difference"). When the sample is limited to compliers, mean off-peak usage significantly differs between households in the price and control groups.

Table (2) also reports home ownership, household income, square footage, and the age of a home. The sample varies across survey questions; notably many households failed to answer questions about the square footage and the age of home. In the initial sample, the average home is 1600

square feet in size and 54 years old. Approximately 77 percent of households own their home and median annual income is between $70,000 and $75,000, indicating that the sample in this study is wealthier than both the national average and the overall population served by the utility. In both the initial and final sample, demographic and household characteristics are balanced between each treatment and control group, with one exception. For both the initial and final sample, households in the price treatment tend to live in larger homes (t=2.10).

To formally test whether the randomization achieved a balance in observables across households, we estimate a linear probability model in which we regress each treatment indicator on the only pre-existing observable variables available for every household in our sample: mean off-peak electricity usage and rate class (flat rate versus time-of-use rate). While the surveys allow us to obtain much more detailed information on household characteristics, we do not use the survey data in our randomization checks, since survey compliance was not 100 percent and their inclusion would confound the interpretation.[23] We discuss survey compliance below, since our analysis of potential mechanisms explaining the treatment effect relies on these data.

The columns labeled "Initial Group" in Table (3) show results, where the sample is comprised of the control group and the price group in column 1, and the control group and price+IHD group in column 2. Neither electricity usage nor rate class is statistically significant in explaining assignment to the initial price and price+IHD treatments. These results indicate that households in each group are balanced on observables, providing evidence that households were randomly assigned to treatments. Still, to control for potential unobservable differences across households in the empirical specifications, we include household fixed effects.

Some attrition occurred in each group, but more so for the price+IHD group, simply because households in this treatment needed to successfully schedule and complete an installation appointment. We investigate the extent to which non-compliance is systematic, since one can imagine that the same factors determining attrition may also impact response to treatment. For example, consider

[23]We initially estimated a linear probability model that included survey demographics as regressors. Though it did not raise serious concerns about the randomization, we choose not to present it due to the impossibility of separating non-random assignment from survey attrition when attempting to interpret the estimates.

households with no one home during working hours. These households will have more difficulty scheduling an installation appointment, and may also be less likely to respond to price increases that occur during working hours.[24] If this is the case, then our estimated effect of the treatment may partly reflect that the households most capable of responding to price events are more likely to be in the price+IHD treatment.

We test for asymmetric attrition by estimating a linear probability model that regresses an indicator variable set equal to 1 for compliers and 0 otherwise on off-peak electricity usage and rate class (for each group separately). The coefficient estimates will be significant if attrition is systematically correlated with the observable regressors. Results are presented in columns 3 and 4 of Table (3). In column 3 the sample is comprised of households initially assigned to the price treatment and in column 4 the sample is restricted to households initially assigned to price+IHD. The significant coefficient on rate class for price+IHD households suggests that some selected attrition may be occurring. Again, household fixed effects will strip out any time-invariant unobservables (such as rate class). Still, in our analysis we use intent-to-treat (ITT) and treatment-on-the-treated (ToT) estimators to account for asymmetries in non-compliance. Overall, we feel comfortable with the integrity of the sample and its ability to support causal inference.

While this research was designed to estimate the information gradient with respect to price elasticity, it is of interest both from an economic and policy perspective to understand the mechanisms driving this effect. Survey responses provide a starting point for this analysis. The general strategy that we pursue is to interact survey responses with treatment indicators in the baseline ITT specifications. However, this is imperfect. First, the survey responses can themselves be interpreted as outcome variables. Second, nearly all households complete the pre-survey, but there is substantial non-compliance across all treatment groups in completing the post-survey. To explore possible asymmetries in survey compliance, Table (4) presents results from regressing pre- and post-survey compliance indicators on observables. With one exception, observables are not significant in explaining survey non-compliance. Post-survey compliers in the price treatment group have lower mean off-peak kWh usage than survey non-compliers. In our discussion of mechanisms, we later

[24]In our study, installation appointments occurred Monday-Friday. No technologies were installed on the weekends.

return to the issue of survey non-compliance and how it impacts the interpretation of our results.

4.2 Raw Data and Mean Differences

Figures (1)-(6) plot raw mean hourly usage by treatment group on each of the six event days. In these figures, mean 15-minute interval consumption is averaged across all households in each treatment group. The shaded area marks the period during which a pricing event occurred. In Figure (1), for example, a DA pricing event was held between 12 pm and 4 pm on July 21.[25] The commonalities between these events are evident: households exposed to information feedback exhibit visibly lower usage during treatment events than price-only and control households. In the hours preceding an event, hourly electricity usage in the three groups is approximately the same, though this appears to change over time.[26] Once the event begins, we observe a divergence in usage between the price+IHD treatment and the other two groups. Households in the price+IHD treatment use less electricity (compared to the control group) in each hour of the pricing event. Interestingly, we do not observe raw differences in hourly usage between price-only and control households, despite households in the former group facing higher prices. Clear temporal patterns also emerge from the plots. These are predictable features of electricity demand which exhibits a high degree of correlation across households due to such things as weather and the commonality of lifestyle patterns.[27]

The high-frequency meter data allow us to retrieve various quantitative estimates of the treatment effect. The coarsest of these is the simple comparison of means in the raw data. Table (5) presents means of electricity usage during pricing events, along with simple differences between each treat-

[25]Recall that the only price change to which any household is exposed during this period is the experimental price change. All households are charged a time-invariant non-experimental price between the hours of 12pm and 8pm, including households on TOU (for whom peak hours are 12pm to 8pm as well).

[26]There is visual evidence of potential habit formation affecting non-event-hour usage, as seen in Figures (3), (5) and (6). Similarly, one may be curious about load shifting. Aside from the pre-treatment hour for the price-only group in Figure (2) and the post-treatment hour for the price+IHD group in Figure (4), the plots suggest that behavioral change is consistent with electricity conservation. In the section on Dynamic Effects, we empirically test for habit formation as well as load shifting.

[27]This is well known and was identified clearly in our power calculations. For a given minimum distinguishable effect, our sample size requirements were greatly reduced by increasing the signal-to-noise ratio through the use of non-parametric controls.

ment group and control using the balanced and unbalanced samples.[28] These comparisons provide support for our hypothesis that there is a positive interaction between information feedback and price response. For each event type, the control group exhibits higher mean usage than either treatment group during high-price hours, and this difference is greatest between the control and price+IHD group. The magnitude of these raw treatment effects varies by treatment and by event type. Response to DA events is higher than to TM events, providing our first evidence that advance notice facilitates a stronger response even under a weaker price incentive. Price+IHD households exhibit between a 12-18 percent treatment effect reduction to all events, as compared to a 0-13 percent reduction by price-only households.

5 Results: Information and Price Elasticity

In the preferred empirical specifications we estimate a difference-in-differences model in which we condition out time-invariant household characteristics and aggregate time shocks

$$k_{it} = \beta_0 + \sum_{g \in \{P, P+I\}} \beta_g D_{it}^g + \gamma_i + \delta_{hd} + \mu_{it} \tag{3}$$

where the dependent variable is the natural log of energy usage by household i in 15 minute interval t. The explanatory variables of interest are the treatment group indicators, D_{it}^g, which are equal to 1 if household i is in group g, and if a pricing event occurs for i in interval t. Controls include household fixed effects (γ_i) and hour-by-calendar date dummies (δ_{hd}). As such, the coefficients of interest are being identified off of variation within households over time. In other specifications, we estimate a simple difference-in-differences model, only include hour-by-calendar date dummies or only include household fixed effects. With respect to inference, it is well known now that serial

[28] Technical issues associated with back-end system software precluded UI from retrieving meter data for some households until after the experiment had begun. Roughly one-quarter of participants are absent from the billing data for at least one event. We are not concerned about the effect of these omissions due to both the exogenous nature of the software issue and the fact that results are quantitatively identical irrespective of the inclusion or exclusion of these households. To make maximum use of the variation in our data, we use the unbalanced panel (as defined by presence in the data for *at least one* treatment event) in what follows. However, in Table (5) and when we present results from the main treatment effect specifications (Tables 6-9), we include the balanced sample as well.

correlation in the dependent variable will, if not corrected, bias standard errors. We address this concern by following the approach that is now standard in the literature[29]; estimates of standard errors from equation 3 are clustered at the household level.

Tables (6)-(9) present the intent-to-treat (ITT) estimates and treatment effect on treated (ToT) estimates, respectively for the unbalanced and balanced subsamples. In each table we begin by estimating a simple difference-in-differences model that does little to condition out systematic confounding variation in the usage data, including only treatment group and price event-period dummy variables (in place of γ_i and δ_{hd} from equation 3). Results from the baseline difference-in-differences approach are reported in Column 1. We sequentially explore the inclusion of more granular non-parametric controls, which (due to the randomization) ought to have the effect of reducing noise and thus increasing precision of the parameter estimates. To the extent that we observe consistency in the point estimates of β_P and β_{PI}, this signifies a robust causal treatment effect. Column 2 presents results with hour-by-calendar date dummies; Column 3 includes household fixed effects but excludes hour-by-date fixed effects; and Column 4 presents results from our preferred specification, which includes both household fixed effects and hour-by-date effects (ie. equation 3). Columns 5 and 6 split out the preferred-specification results by event type, with Columns 5 and 6 reflecting DA and TM events, respectively.

5.1 Intent-to-Treat

In the ITT estimator, D_{it}^{P} and D_{it}^{P+I} in equation (3) denote initial assignment to the price and price+IHD treatments, irrespective of whether the household dropped out or was removed from the study. The coefficients, $\hat{\beta}_P$ and $\hat{\beta}_{PI}$, in this specification are consistent estimates of the average percentage change in electricity usage from assignment (though not necessarily compliance) to the price and price+IHD treatments during pricing events. Results from the ITT regression are relevant due to the well-documented difficulties encountered by electric utilities in engaging customers in their various initiatives. They provide utilities and their regulators a measure of the overall benefits

[29] See Bertrand et al. 2004.

available were they to implement pricing events (with or without information displays).

Table (6), which reports results using the full (unbalanced) sample of usage data, highlights that households with real time information feedback are significantly more responsive to price changes than those without. When DA and TM events are pooled together (columns 1-4), households with an IHD reduce usage by 11 to 14 percent and are over three standard deviations more responsive to pricing events as those without IHDs (column 4). This treatment differential is robust to the controls employed. Coefficient point estimates on the price+IHD group and the 8 to 10 percentage point treatment differential between households in price and price+IHD groups are both significant with over 90 percent confidence. These results provide strong evidence that the cumulative effect of real-time information feedback can meaningfully increase the price elasticity of demand.

In the absence of information, the price effects are weak. Price-only households on average decrease usage by 2 to 6 percent during pricing events. Even using the full array of controls, when pricing events are pooled together the response to pricing events is not statistically significant. These results confirm the pattern illustrated in the daily plots which suggests little if any response to pricing events.

A comparison of the treatment effect across columns 1-4 reveals a treatment differential that is robust both in terms of magnitude and statistical significance to the controls included. In the simple difference-in-differences, the treatment differential is 8 percentage points and is significant with over 90 percent confidence. That the inclusion of household and time controls does not meaningfully alter the magnitude of treatment effects provides further evidence for the integrity of the randomization.

We break out the events separately by type in columns 5 and 6, and continue to find that households with IHDs are significantly more price elastic. The percentage point treatment differential between households in the two treatment groups is robust to event type and significant with over 85 percent confidence. However, individuals are more responsive, both in economic and statistical significance, to pricing events that occur with more advance notice. This is true despite the fact that the price increase in TM events is more than twice that for DA events ($1.25 as compared to $0.50).

Households assigned to the price group reduce usage by 7 percent during DA events where this response is weakly significant. In contrast, they are largely unresponsive to TM events and if anything increase usage during them. We find similar patterns across event type for households assigned to the price+IHD treatment. With advanced notification, these households reduce usage by 17 percent as compared to an 8 percent (but not statistically significant) response with thirty minute notification. Even with strong financial incentives, with only 30 minutes of warning individuals may not have the ability to respond to a price change.

Table (7) is the balanced-panel analog to Table (6). The sample is comprised of households observed during every pricing event, and we find qualitatively and quantitatively indistinguishable results from those reported in Table (6). Columns 1-4 show that households with IHDs are significantly more responsive to pricing events than those without. This 9 to 10 percentage point treatment differential is robust both in magnitude and statistical significance to the inclusion or exclusion of controls. The magnitude of the usage response for households in the price+IHD group ranges between 10 and 13 percent, and is estimated with over 95 percent confidence. Again we find less of a response, both in magnitude and statistical significance, among households assigned to the price only group, where the reduction in usage amounts to 0 to 4 percent. Similar to the results presented in Table (6), households are more responsive to pricing events with more advanced notification (columns 5 and 6).

5.2 Treatment on Treated

The ToT specification identifies the causal effect of the price and price+IHD treatments on compliers. These estimates are most relevant for understanding whether customers that face a sharp reduction in information acquisition costs will be more price elastic. The ToT approach uses initial treatment assignment as an instrument for final treatment status and is estimated using two-stage least squares. The instruments are valid because the randomization ensures that the exclusion restriction holds, and strong due to the high rate of compliance with initial group assignment. In the unbalanced (balanced) panel compliance was 98 (98) percent, 95 (98) percent, and 72 (84) percent

in control, price-only, and price+IHD, respectively.

Estimates of ToT specifications using the full sample and restricted sample of usage data are presented in Tables (8) and (9) respectively. The qualitative pattern mirrors the ITT results, but with an even stronger treatment differential between price-only and price+IHD households. Compared to the ITT estimates, the magnitudes are slightly larger as a result of estimating the percentage change in usage from treatment rather than initial assignment to treatment. Both the treatment differential and standard errors on the coefficient estimates are largely insensitive to the inclusion or exclusion of household and time controls. In our preferred specification, households with IHDs are 13 percentage points (3-4 standard deviations) more responsive to pricing events as those without, where this treatment differential is present with 95 percent confidence. Households in both treatment groups are more responsive to advance notification events. We observe a 7 percent reduction in usage for households in the price only group (where the coefficient estimate is weakly significant when using the full sample of data) and an 18 to 22 percent reduction for households in the price+IHD group. For both types of pricing events, we continue to observe a treatment differential with over 85 percent confidence. Finally, comparing both the coefficient estimates and treatment differentials in Tables (8) and (9) we find similar treatment effects, suggesting that the estimated information gradient is not sensitive to our choice of sample.

Many readers may wish to see these results presented as elasticities, which is of course mathematically possible. The implied arc elasticities are low (less than -0.12). However, it is unclear as to what information is actually conveyed by these statistics due to the magnitude of the price increases (200 percent during CPP events and 600 percent during DR events). The range of the price increases may contain regions of highly elastic demand as well as regions of inelastic demand, allowing for the possibility that lower (absolute value) price changes may induce the same absolute behavioral response. In any case, these estimates are a lower bound for reasons that we will discuss in the section on dynamic effects.

Before turning to dynamic considerations, though, we will examine how the treatment effects correlate to survey responses. There are many explanations as to why the absence of information feedback inhibits customers from fully responding to price changes. These may have quite different

implications for what we learn about consumer choice (in this setting), and how the results should be used by policymakers.

5.3 Secondary Results and Robustness

The RCT was designed to identify causal effects relating to our primary hypothesis about the information gradient in price response. In this section we seek to add context on how the main result should change the way we think about consumer choice and behavior. In particular, we test the robustness of the information treatment differential on subsets of households (by observable characteristics), and also examine whether observables help to explain *how* households are actually achieving demand reductions. To do this, we gathered a rich dataset of household attributes and experiences during the experiment from responses to the pre- and post-surveys. We also rely on event specific data indicating whether a household confirmed receipt of pricing event notification. How the survey responses (and confirmation of event notification) correlate with treatment outcomes offers insight into the mechanisms underpinning our primary result.[30]

Potential explanations for the strong information effect are many. Our preferred hypothesis–that is, the one that we believe is most consistent with observed behavior–is that experience with the displays facilitates *learning* about energy choices and the mapping from usage to expenditure. This mechanism is consistent with a framework in which baseline information about quantity is imperfect, and whereby IHDs serve to inform consumers and (in doing so) influence their subsequent choices. In this section, we first present some evidence that attempts to rule out the leading alternative hypothesis – that IHDs generate more awareness, and consumers who possess them become more price responsive for this reason. We provide evidence that this explanation is not likely to be responsible for the information gradient (though we also show that awareness of price events is correlated in both the price-only and price+IHD groups).

Subsequently in this section, we also explore the practical mechanisms by which households respond

[30]In the specifications to follow, we interact treatment indicators with survey responses that are themselves outcome variables. These should not be interpreted as heterogeneous causal effects. Rather, they should be thought of as cross-tabulations of the treatment effect with various survey responses.

to price changes. This line of inquiry is not intended to explain the nature of the information gradient, but rather to aid our understanding of what households actually *do* to respond. A consumer's choice set and subsequently his ability to respond to prices will likely be related to certain household patterns or attributes. For example, differences in the frequency with which households are at home during the day or have central air conditioning could affect the ability to respond to price incentives. Indeed, we show that these are correlated with consumer response, but that they do not cause the information gradient. Instead, the balance of evidence leads us to believe that consumers learn through experience, and that this plays an important role in the treatment differential.

Due to our reliance on survey data, the sample for the analyses that follow is restricted households that completed the pre- or post-survey (survey "compliers"). As a result, the estimates may be vulnerable to differences in survey non-compliance across groups that are also correlated with unobservables, and should be interpreted with this caveat in mind. But given the high survey response rate, these results provide a reasonable starting point to understand consumer behavior. Our intention is that insights from this section, while they do not answer all of the questions a reader may have, could lead to future research hypotheses that motivate experiments designed to distinguish between subtle behavioral mechanisms.

5.3.1 Notification and Awareness of Price Events

A plausible alternative to the "learning" hypothesis is that IHDs facilitate a heightened awareness of price events. Our experimental design sought to control for this by making electricity price changes salient to all consumers who were exposed to them. This was achieved by having the utility send all customers in the price and price+IHD groups notification in the form of a combination of a text message, email and/or phone call in advance of each event. The messages alerted households that a pricing event was going to occur and informed them what the price would be. Still, we would like to test that there is no additional awareness effect from the IHDs that is driving the main treatment differential.

One indicator that customers were aware of pricing events is if they indeed received notification of them. If for some reason, notifications were not received then the IHD may provide incremental and valuable awareness of them. If we observed which households actually received event notification and which did not, we could examine the correlation between notification confirmation and treatment response. Fortunately, for each pricing event, the utility collected data confirming the receipt of event notification and documenting the extent to which notification information was conveyed if notification was not confirmed. To acknowledge notification of an event, a household needed to a push a button on the telephone (informing customers of an event) at the end of of an automated phone call or a click a button confirming receipt of a notification email. These events as well as households who had notification texted to them were defined as having event confirmation. The remaining events were classified as not confirmed or intermediate, where the latter category includes instances when someone answered the phone but hung up before the automated message ended or an automated message was left on a voice mail.

Table (10) shows that 34 to 45 percent of price-only households and 24 to 43 percent of price+IHD households confirmed receipt of the event notification, and 70 to 80 in either group achieved at least an intermediate confirmation status. We conclude that awareness of the events was high, suggesting that the treatment events were front-of-mind for a majority of households.[31] Were the differential response to exist because IHDs increase awareness of price events, then we would expect to see that, conditional on confirmation/awareness, households in the two groups would respond equivalently. To test this hypothesis, we estimate

$$k_{it} = \beta_0 + \sum_{g \in \{P+I\}} \beta_g D_{it}^g * A_{it} + \gamma_i + \delta_d + \sigma_{hw} + \mu_{it} \tag{4}$$

in which we interact the treatment indicator D_{it}^g with A_{it}, a variable set equal to 1 if receipt of event t notification was confirmed, and a household was thus aware of the event. The overall results ("All events") are presented in columns 1 and 4 of Table (11).

[31]The post-survey also solicited self-reported "awareness" of pricing events while they were occurring. Earlier versions of the paper present statistics on these responses, with 80 to 90 percent of households claiming to have been aware of events. Results are available from the authors upon request.

In columns 1-3 the "intermediate" confirmation and full confirmation indicators are interacted separately with the treatment dummy. We first notice that the treatment effect for price+IHD households is similar between the intermediate and confirmed designations. This motivates the specifications presented in columns 4-6, in which the intermediate confirmations are defined as fully confirmed. It is clear that event notifications are important, as households confirming receipt of them exhibit a larger response overall and for the DA events.

In columns 4-6, we see that confirmation of event notification alone cannot the treatment differential between those with and without IHDs.[32] Conditional on confirmation of event notification, we reject the null that the coefficient estimates are equal with 95 percent confidence overall, and with 90 and 85 percent confidence when estimating DA and TM events separately. At the same time, conditional on no confirmation we find a statistically indistinguishable response across the price and price+IHD groups. This suggests that IHDs do not appear to be either informing households of events or enabling unaware households to respond. Together, these results suggest that IHDs trigger or facilitate a mechanism beyond price event awareness.

5.3.2 Home During Daytime

All of the price events occurred at times that overlap with the traditional workday. Households with daily schedules that find at least one person at home during these hours will have a broader range of activities with which to respond to price events. At the same time, compliance appears to be correlated with this household attribute. To remain in the study, households assigned to the information treatment needed to successfully schedule and install an IHD. Households who are not at home are less likely to have an IHD installed, and also may be less responsive to events due to the fact that the house is empty when prices are high. While the interpretations of ITT and ToT estimators are robust to this possibility (thus maintaining legitimacy of our claim to internally validity), we are still interested to see how the response to pricing events varies by how frequently an individual is at home. To do this, we rely on pre-survey data that asked households if someone

[32]It is clear that awareness of events is important as households in either treatment group who confirm receipt of event notification exhibit a larger response to pricing events.

was generally at home during the day, after school or after work, and create indicator variables that we interact with assignment to treatment.[33] Given that almost all households completed the pre-survey, including those who failed to have IHDs installed, the first specification in which the at-home dummy variables are interacted with treatment assignment are very similar to an ITT estimator. We also present the ToT analog.

The first two columns of Table (12) show that the information feedback group has fewer households with someone home during the day than the price-only group, both before and after accounting for attrition. Slightly over 50 percent of the households assigned to the price-only group have someone home during the day, as compared to 40 percent for the price+IHD treatment. Interestingly, when we restrict the sample to compliers this pattern remains, implying that we should not be too concerned about installation-related attrition being closely correlated to participants' schedules. Conditional on completing the pilot, 44 percent of households in the feedback treatment and 52 percent of households in the price treatment tend to have someone home during the day. If households with someone home during the day are more responsive to pricing events, then the differential response attributable to information feedback will be, if anything, attenuated.

As such, having someone present in the household is not driving the differential effect between those with and without feedback. Conditional on someone tending to be home during the day, households with an IHD are more responsive to price changes, where the difference in treatment effects ranges from 8 to 12 percentage points in the ITT specifications. The treatment differential is larger in the ToT specifications, ranging between 9 and 17 percentage points, and distinct with roughly 90 percent confidence for day-ahead pricing events.

Within each treatment cell we observe some interesting differences in the response to pricing events based on household schedule. As expected, households that tend to have someone at home during the day are more responsive (as compared to households with someone home after work) to the short-notice TM pricing events. Households with someone home after school appear to be similar

[33]This survey response pre-dates treatment assignment and is thus predetermined with respect to the treatment. However, it is also only an approximate measure of the probability with which households were home during the day during treatment events.

to those home during the day for these events, though "after school" may have a different meaning in different households. These results suggest that short notification events inhibit households that are not at home during the day from responding to them.

5.3.3 Central Air Conditioning

We investigate correlations with respect to central air conditioning for two reasons. First, an unfortunate sampling draw could have yielded unobservable (at the time of randomization) differences between the groups along this potentially important dimension. We would like to rule out the possibility that a narrative along these lines explains the primary result of the information-treatment gradient. Also, given the importance of central air conditioning as part of the home-energy portfolio, we want to understand the extent to which consumers are using it to respond to price. It is a high usage activity that can easily account for large usage reductions when turned off (or down) during events.

Table (13) shows the percentage of households with central AC and reports treatment effects for households with and without it. Results are contrary to our initial expectations. Households *without* central AC in both treatment groups are on average more responsive to pricing events. The first implication of this result is that central AC cannot be a necessary or sufficient condition for the information gradient to be present. Second, this implies that households exhibiting large responses are doing so by adjusting along other, likely smaller, margins. We view this as a reflection of some underlying selection process whereby non-AC household members are more price-responsive. For instance, younger households being more interested in energy conservation, but may also live in dwellings less likely to have central AC (as opposed to some feature of having central AC that somehow inhibits conservation during events).

5.3.4 Experience

We have shown that three potential mechanisms – awareness, daily schedule, and central air conditioning – do not account for the information gradient in price response. We now hypothesize

that IHDs facilitate learning about the electricity usage (and thus electricity expenditure) associated with the portfolio of household production alternatives. To the extent that customers observe and interact with the IHD, they are gaining information about features of the mapping between electricity-consuming actions and the usage that they require. For example, if a consumer views the display before and after turning on her air conditioner, she will notice that it consumes a large amount of electricity. This knowledge, if accumulated analogously for multiple appliances, may better equip households to respond to price changes. As evidenced by the economic and statistically significant treatment effects for households in the price-only group, even without IHDs, households are aware of some of the margins of adjustment available to them. However, they may not fully understand the full choice set of energy saving behaviors, and the quantity of electricity usage associated with these.

To test if frequent experience with the IHD increases price responsiveness, we use responses from the post-survey. Households were asked, "How many times per week did you look at the IHD in the first month that it was activated?"[34] Table (14) shows that most households frequently experimented with their IHD in the first month that they received it. Conditional on answering this question, only 6 percent of respondents did not look at it and approximately 65 percent of respondents looked at it more than 5 times. We conclude that households with an IHD are actively engaging with it.

In columns 2-4, we present results from the estimation of equation 4, where A_i is now a vector of indicator variables describing how frequently a household looks at the IHD. Households who engage most frequently with the IHD are significantly more responsive to price changes than others.[35,36]

[34]Note that IHDs were installed two to four months before the first pricing event occurred, leaving ample time for households to interact with them before the treatment events occurred.

[35]It is possible that the direction of causation is such that those inherently more responsive to price also look at the IHDs more frequently. However, if this more engaged cohort exists, then we would expect them to be present with similar frequency in the price-only group. Given that we observe a differential response across those with and without IHDs, there is little evidence to support a hypothesis about reverse causality.

[36]Households reporting to never look at the displays have large measured responses as well. However, these are few in number and there are several reasonable explanations. It is possible that the respondent is not the person responsible for the energy decisions that leads to the large measured treatment effects. Also, perhaps instead of using the IHD, these households log into the web portal, or already were familiar with all the margins of adjustment available to them (whereby the IHD provides no additional value). Or, when notified of pricing events, these households may simply leave the premise for the duration of the price

This suggests that more frequent experience with the IHDs facilitates learning, perhaps about the quantity of electricity consumption. This is consistent with households gaining a fuller understanding about what is discretionary and what is not, or about the mapping of behavior into electricity usage.

6 Dynamic Effects

Despite confining the pricing events to small windows in time, it is possible (indeed, plausible) that the behavioral response that is induced may affect consumer choice in both the short run and the longer run. In this section we explore these two views of dynamic household usage which are important quantitatively and also in helping to fill in the narrative about consumer behavior. Quantitatively, insofar as households are conserving outside of the treatment window, this would attenuate our estimates of the treatment effect by lowering the baseline for within-household comparison. In the short run, there are competing hypotheses. Usage in hours on either side of the event window may increase, relative to normal, which is consistent with households pre-cooling their homes before anticipated price increases or delaying activities (e.g. laundry) until the price returns to its normal level. Alternatively, activities undertaken in response to high prices may spill over into the hours preceding or following the events. If true, this would manifest as lower usage during those times.

Table (15) reports estimates from the ITT specification from equation 3, but with the addition of indicator variables for the two hours preceding and following price events for each treatment group. The results reveal that conservation during pricing events spills over in to adjacent hours for the price+IHD group, but not for the price-only group. The findings clearly do not support the so-called "pre-cooling" hypothesis. Rather, in the hours preceding an event, households exposed to day-ahead notification exhibit no change in usage while those with feedback significantly reduce usage by 9%. We also find that with advance notification the treatment effects spill over into the two hours following an event. In contrast, we find no significant evidence of spillovers (or load

increase.

shifting) in response to TM events, probably because of the limited amount of time to anticipate them. For neither event type do we find quantitative evidence of load being shifted to the hours immediately preceding or following events. The direction of this short-run affect has environmental implications, since usage reductions in response to treatment are not significantly offset by increases in usage at other times.

In the medium-term (weeks to months) as households are exposed to multiple events their usage decreases in meaningful ways during *non-event* days. We call this effect "habit formation", and it is responsible for large conservation effects on non-event days over the months of our study. Table (16) reports coefficients from a single regression in which we estimate separate trends for each hour of the peak period over the 62 days of July and August.[37] For example, row 1 displays the trend in kWh for the period noon to 1pm, implying an average daily decrease (gradient) in usage of 0.19 percent for price-only households during this noontime hour. This corresponds to a 12 percent decrease in usage from July 1 to August 31. At the upper end of the coefficient range conservation effects are equivalent to 17 percent during 7-8pm for the price+IHD group. We also find that the conservation trend is steeper for price+IHD households in some hours, but not all. The maximum differential between households with and without feedback occurs between 4-5pm and translates into 3 percent (15 versus 12) more conservation over these months.

Given the magnitude of the treatment differential during the price events, it is noteworthy that habit formation (across the two groups) throughout non-event days is quite similar. This dichotomy suggests, perhaps, that recurring price events induce a response irrespective of the presence of information feedback; but that the information enables a more sophisticated and precise response to short-run incentives. If true, this lends further support to the hypothesis that a different sort of learning is taking place when facilitated by IHDs.

Another important implication of the dynamic effects relates to the econometric identification of the baseline treatment effect, which comes from within-household variation. Reductions in usage

[37] We believe that panel balance issues may be relevant for this analysis. We therefore use the balanced panel sample in Table (16) as a conservative approach, foregoing some variation in the data in favor of including households that are present throughout the entire set of treatment events.

outside of the event windows will lower the baseline against which treatment-period reductions are measured, biasing the primary treatment effects towards zero. Further, since spillovers are stronger for price+IHD households, the treatment *differential* will also be attenuated, implying an even larger gradient of information feedback on price elasticity. Finally, all of these conservation effects imply a favorable result with respect to greenhouse gas emissions.

7 Welfare and Policy Implications

The assumption of full information underpins countless important economic results about market efficiency, including any application of the first and second fundamental theorems of welfare.[38] It is not surprising, then, that the absence of full information may lead to inefficient market outcomes. The main result of this paper points to the importance of information, whereby consumers with additional information about the quantity of consumption make vastly different decisions than their uninformed counterparts. This finding not only contributes to the way we think about consumer choice in the electricity setting, but may also extend to other settings in which quantity is shrouded from the consumer, such as cellular phone minutes, Internet or data, water, heating oil and natural gas usage.

However, we confine our welfare discussion to the electricity sector, which exhibits a particularly troublesome inefficiency: the disconnect between wholesale and retail prices.[39] Time-varying electricity prices that better reflect generation costs would mitigate short-run inefficiencies by aligning social cost with marginal benefit, reduce over-investment in capital in the long run, lessen the harm from market power in the generation sector, and (depending on the location and timing of consumption) potentially reduce the external cost of emissions (Borenstein 2002, Borenstein 2005, Borenstein and Holland 2005, Holland and Mansur 2008).[40] A historical barrier to the implementation of time-varying retail prices was the ability to collect high-frequency usage data and transmit

[38] See Greenwald and Stiglitz (1986).

[39] There are also others, most notably market power and external costs of pollution.

[40] Other long-run benefits include a reduction in ramping costs, as well as potential savings in transmission and distribution infrastructure.

this information to customers (Joskow and Wolfram 2012). The device used in our study removes this barrier, and is precisely the type of technology being considered by regulators to facilitate dynamic retail electricity pricing. Our results highlight the potential of time-variant pricing to reduce usage, especially when coupled with information feedback.

Now that the technology exists to implement time-variant pricing, electric utilities and their regulators are trying to understand whether to deploy HAN technology in the field, and if so, the appropriate payment structure and nature of deployment. In this section, we first estimate the expected net private benefit to consumers of purchasing an IHD, and then shift our focus to estimating net social benefits. The magnitude of net benefits accruing to households provides insight into the likelihood of a successful free-market solution in which households voluntarily purchase HAN technology for their homes.

Assuming that all households are placed on a time-variant tariff (similar to the one we implemented), we now use our treatment effect on treated households to quantify the potential savings to households of owning an IHD. Table (17) shows the change in expenditure from price increases of either \$0.50/kWh or \$1.25/kWh with and without an IHD. The difference between the Price and Price+IHD columns represents the incremental value of an IHD. We vary the number of hours - 20 hours, 40 hours and 60 hours - that households are exposed to price increases. If a household is exposed to 40 hours of \$0.50/kWh or \$1.25/kWh prices respectively, the IHD would facilitate \$3-\$8 in savings for the household. Note that this is an overestimate of the marginal net benefit of utilizing the IHD since we omit from our measure of cost any disutility incurred by households from altering their behavior. Currently, a basic IHD is available at retail stores for well under \$40,[41] though many more expensive models exist (including the ones used in this study). According to our estimates, at a 5 percent discount rate, the payback period of a \$40 device is over 20 years for \$0.50/kWh events, but under 6 years for \$1.25/kWh events. Even if these payback periods were not underestimates, it seems unlikely that the "average" electricity consumer would find the voluntary purchase of an IHD an attractive investment.

Since many of the benefits of energy conservation are external to the consumer, we also examine

[41] e.g. the "Kill A Watt", which at the time of this draft is available for \$28.97 at a major online retailer.

the net social benefit of deployment, accounting for some of the general equilibrium effects of IHDs and dynamic pricing on the need for excess generation capacity. This estimate will be informative as to the role of regulation and government intervention in HAN deployment. However, the net social welfare gains are more difficult to quantify. Borenstein and Holland (2005) estimate that the industry-wide deadweight loss (excluding external environmental costs) from flat rate electricity pricing is on the order of 5-10 percent of the wholesale electricity market. In 2009, when $350 billion of electricity was sold in the United States, this translates into $17-$35 billion dollars, a portion of which is attributable to the residential sector. The question is, how much of this deadweight loss would be remedied by the presence of both CPP (or a similar dynamic pricing scheme) and feedback technology (IHDs)?

To get a sense for this, we apply the price elasticities estimated in our setting to a previously developed simulation evaluating the long-run welfare gains from switching one-third of residential customers in the U.S. onto RTP (Borenstein 2005). While this is a different flavor of dynamic pricing than the one we implement, a well-executed CPP program targeting the most grid-strained days will likely yield a majority of the benefits that a RTP scheme would. Using the price elasticities in our study, we find that placing one-third of customers on RTP would increase annual total surplus by $200 to 250 million (in year 2000 dollars).[42] Approximately 35 to 50 percent of these gains would accrue to residential customers remaining on the flat-rate tariff, since they now face lower average electricity prices because of the efficiency gains. To deploy IHDs to one-third of households nationwide would cost approximately $1.4 billion dollars (35 million households * $40/IHD), implying a payback period of between 7 and 9 years.[43] This payback period is an upper-bound estimate since the simulations in Borenstein (about which he is perfectly transparent) do not consider market power, ramping costs, and transmission and distribution costs. There may also be economies of scale in the procurement of the feedback technology, reducing the up-front cost of deployment. Finally, since the exhibited responses to feedback are conservationist as opposed to load shifting, there will be environmental gains from a decrease in generation-related pollution.

[42]This turns out to be true in the simulations when starting from a very low base elasticity (-0.050) or a somewhat higher one (-0.100).

[43]The payback period assumes a discount rate of 3 percent, and annual gains of either $200 million (9 years) or $250 million (just over 7 years).

While our estimates are internally valid, our sample is select in that it is comprised of households with paperless billing and broadband that volunteered to participate in this study. Our results must be interpreted in this context. That said, the main qualitative result may be broadly applicable to other locations and settings, most clearly those in which choices ought to incorporate a metered quantity (e.g. cellular phone, water, or cellular data usage).

To help gauge the external validity of our results and the potential of a national deployment of HAN technology, we estimate the ITT and ToT specifications, weighting our observations to reflect the income distribution in the United States. Consistent with our setting, we assume that this deployment will target technologically advanced households, specifically those with broadband Internet. This is quite a large proportion of the U.S. population; over 77 percent had Internet access in 2009 and over 27 percent of the population has broadband.[44] Results are shown in Table (A-1). They are qualitatively similar to the preferred baseline estimates, and if anything show a stronger treatment effect in IHD households.

8 Conclusion

This paper shows that information feedback about quantity of consumption increases the price elasticity of demand, using the residential electricity sector as the field setting. The experimental design allows us to identify both the price elasticity as well as the marginal contribution of information feedback. Households only experiencing price increases respond to them, reducing usage by 0 to 7 percent. But those exposed to both price increases and information feedback produce a 8 to 22 percent reduction. The incremental contribution of feedback is both economically and statistically meaningful, inducing a three standard deviation increase in the behavioral response. By industry standards, the magnitude of these effects is remarkable and points to a level of customer engagement not often seen.

The usage reductions extend beyond the pricing event windows, to both non-event hours on event

[44]International Telecommunications Union.

days and non-event days. For households in the price+IHD group, energy conservation spills over into the hours immediately preceding and following events. The trend over July and August suggests that households in both treatment groups are forming conservation habits. The combined effects of spillovers and conservation habit formation cause us to underestimate our primary treatment effects and the information gradient on price elasticity.

The totality of evidence lends support to the hypothesis that information feedback facilitates learning. First, customers who more frequently look at their IHD are significantly more responsive to pricing events. It is natural to expect that frequent experience with these displays may lead customers to learn about electricity usage, particularly the relationship between electricity services and their cost. Second, we find that informed households are achieving conservation in ways not confined to central AC, the most energy-intensive use. In fact, IHD households without central AC are on average more responsive to price changes that those with it. This is consistent with a story by which information feedback allows people to learn about many available margins of adjustment. Lastly, we observe some interesting nuances in the nature of learning by noticing the *similarity* of habit formation between the price-only and price+IHD groups but the *difference* in event-period behavior. Information feedback appears to provide consumers with a more precise ability to respond to rapid price changes, but the price treatment alone appears to cause the development of conservation habits.

These results confirm the practical importance of one of economics' most ubiquitous assumptions – that decision-makers have perfect information. Indeed, the absence of perfect information is likely to cause substantial efficiency losses both in this setting and others in which quantity is also infrequently or partially observed by decision-makers. Under conservative assumptions, our results imply that a widespread deployment of information feedback technology would pay for itself in no more than 7-9 years in the residential electricity setting. An additional social benefit of feedback comes in the form of greenhouse gas abatement, with IHDs inducing up to a 3 percent conservation effect in the medium-run. Our study suggests price remains a promising lever by which to influence electricity demand when consumers are well-informed about their own usage.

References

[1] Allcott, H. (2011a). "Rethinking Real-Time Electricity Pricing," *Resource and Energy Economics*, 33(4): 820-842.

[2] Allcott, H. (2011b). "Social Norms and Energy Conservation," *Journal of Public Economics*, 95(9-10): 1082-1095.

[3] Bertrand, M., E. Duflo, and S. Mullainathan (2004). "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119(1): 249-275.

[4] Borenstein, S. (2002). "The Trouble with Electricity Markets: Understanding California's Restructuring Disaster," *The Journal of Economic Perspectives*, 16(1): 191-211.

[5] Borenstein, S. (2005). "The Long-run Efficiency of Real-time of Electricity Pricing," *The Energy Journal*, 26(3): 93-116.

[6] Borenstein, S. and S. Holland (2005). "On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices," *Rand Journal of Economics*, 36(3): 469-493.

[7] Chetty, R., A. Looney, and K. Kroft (2009). "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99(4): 1145-1177.

[8] DellaVigna, S. and J. Pollet (2008). "Investor Inattention and Friday Earnings Announcements," *Journal of Finance*, 64: 709-749.

[9] DellaVigna, S. (2008). "Psychology and Economics: Evidence from The Field," *Journal of Economic Literature*, 47: 315-372.

[10] Faruqui, A. and S. Sergici (2010). "Household Response to Dynamic Pricing of Electricity: A Survey of the Experimental Evidence," *Journal of Regulatory Economics*, 38: 193-225.

[11] Faruqui, A., S. Sergici, and L. Akaba (2012). "Dynamic Pricing in a Moderate Climate: New Evidence from Connecticut," Working Paper.

[12] Feng, L. and M. Seasholes (2005). "Do Investor Sophistication and Trading Experience Eliminate Behavioral Biases in Financial Markets?" *Review of Finance*, 9: 305-351.

[13] Greenwald, B. and J. Stiglitz (1986). "Externalities in Economies with Imperfect Information and Incomplete Markets," *Quarterly Journal of Economics*, 101 (2): 229-264.

[14] Grubb, M. and M. Osborne (2012). "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock," Working Paper.

[15] Harding, M. and D. Rapson (2012). "Does Absolution Promote Sin? Green Marketing and Energy Demand," Working Paper.

[16] Herter Energy Research Solution (2012). "SMUD's Residential Summer Solutions Study," Working Paper.

[17] Hirshleifer, D., S. Lim and S. H. Teoh (2009). Driven to Distraction: Events and Under-reaction to Earnings News, *Journal of Finance*, 64(5): 2289-2325.

[18] Hossain, T. and J. Morgan (2006). ". . . Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay," *B.E. Journals in Economic Analysis and Policy: Advances in Economic Analysis and Policy*, 6(2): 1-27.

[19] Holland, S. and E. Mansur (2008). "Is Real-Time Pricing Green? The Environmental Impacts of Electricity Demand Variance," *The Review of Economics and Statistics*, 90(3): 550-561.

[20] Ito, K. (2011). "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing," Working Paper.

[21] Jacobsen, G., M. Kotchen, and M. Vandenbergh (2010) "The Behavioral Response to Voluntary Provision of an Environmental Public Good: Evidence from Residential Electricity Demand," *European Economic Review*, 56: 946-960.

[22] Joskow, P and Wolfram, C. (2012) "Dynamic Pricing of Electricity," forthcoming, *American Economic Review, Papers and Proceedings.*

[23] Lacetera, N., D. Pope, and J. Sydnor (2012). "Heuristic Thinking and Limited Attention in the Car Market," forthcoming, *American Economic Review.*

[24] List, J., (2003). "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics*, 118(1): 41-71.

[25] Reiss, P. and M. White (2005). "Household Electricity Demand Revisited," *The Review of Economic Studies*, 72: 853-883.

[26] Simon, H. (1955), "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69(1): 99-118.

Figures and Tables

Figure 1: July 21, 2011: 4hr $0.50 increase, day-ahead notice

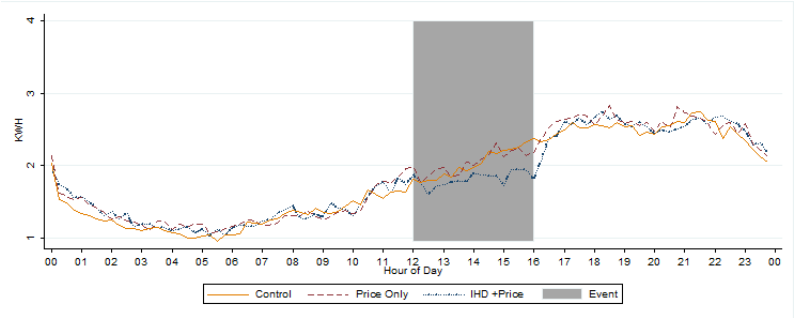


Figure 2: July 22, 2011: 4hr $1.25 increase, 30-min notice

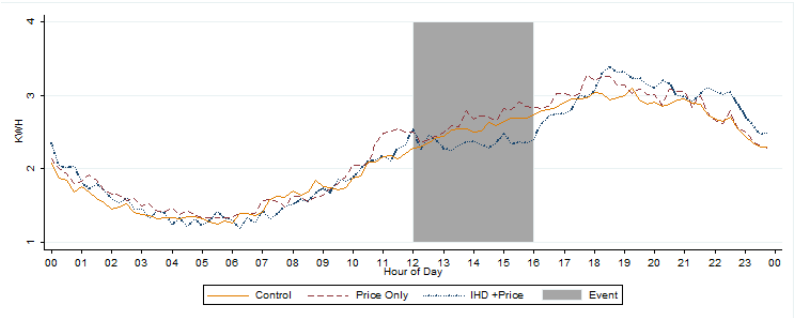


Figure 3: August 4, 2011: 2hr $0.50 increase, day-ahead notice

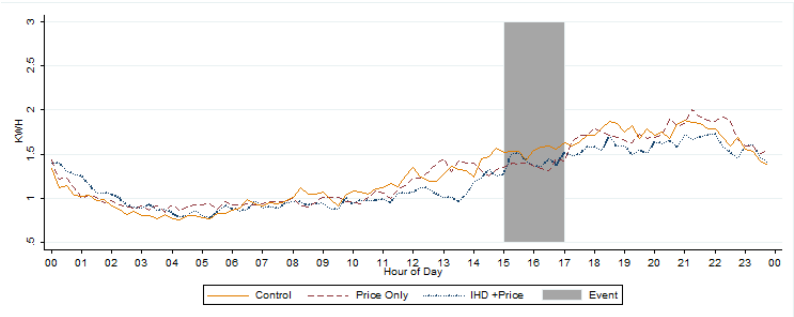


Figure 4: August 10, 2011: 2hr $1.25 increase, 30-min notice

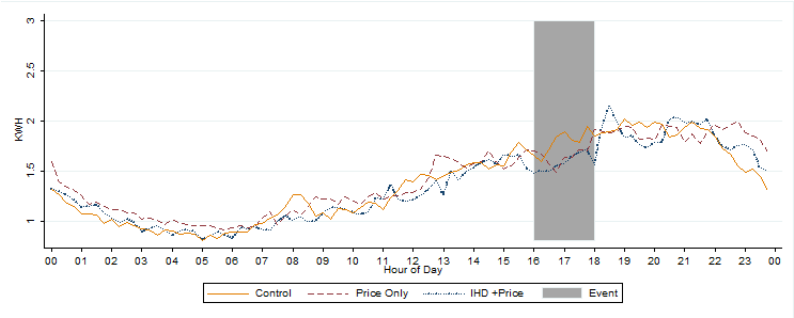


Figure 5: August 17, 2011: 2hr $1.25 increase, 30-min notice

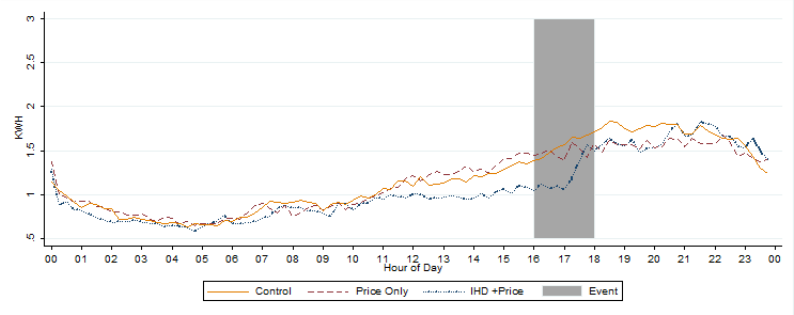


Figure 6: August 26, 2011: 4hr $0.50 increase, day-ahead notice

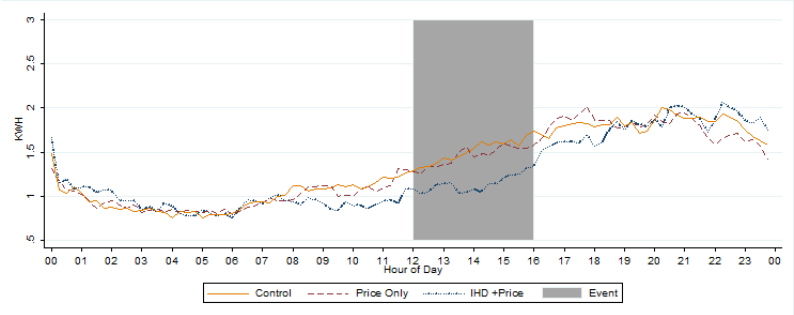


Table 1: Treatment Events

Event Date	Desc	Type	Start Hour	High Temp	Mean Temp	Humidity
07/21/11	4 hr $0.50	DA	12	89	82	75
07/22/11	4 hr $1.25	TM	12	103	90	61
08/04/11	2 hr $0.50	DA	15	80	74	68
08/10/11	2 hr $1.25	TM	16	88	80	63
08/17/11	2 hr $1.25	TM	16	86	75	64
08/26/11	4 hr $0.50	DA	12	84	78	69

Table 2: Summary Statistics by Control and Treatment Group

Panel A: Initial Group

	Control		Price			Price+IHD		
	Mean	Obs	Mean	Obs	Difference	Mean	Obs	Difference
Off-peak usage (kWh/h)	1.159	207	1.279	130	0.121	1.203	100	0.044
	(0.687)		(0.737)		(1.524)	(0.646)		(0.542)
Peak usage (kWh/h)	1.422	207	1.529	130	0.107	1.383	100	-0.038
	(1.107)		(1.034)		(0.887)	(0.954)		(-0.298)
TOU Rate (1=yes)	0.184	207	0.200	130	0.016	0.240	100	0.056
	(0.388)		(0.402)		(0.373)	(0.429)		(1.153)
Home ownership (1=yes)	0.768	203	0.798	129	0.030	0.773	97	0.005
	(0.423)		(0.403)		(0.641)	(0.42)		(0.091)
Annual income ($1000)	72.00	203	74.00	129	2.000	71.00	97	-0.001
	(29.00)		(29.00)		(0.690)	(31.00)		(-0.181)
Home size (1000 square feet)	1.529	189	1.880	119	0.351 **	1.451	91	-0.078
	(1.11)		(1.83)		(2.100)	(1.14)		(-0.550)
Age of home (years)	52.423	156	57.619	97	5.195	52.239	71	-0.184
	(30.29)		(31.34)		(1.309)	(26.94)		(-0.044)

Panel B: Final Group

	Control		Price			Price+IHD		
	Mean	Obs	Mean	Obs	Difference	Mean	Obs	Difference
Off-peak usage (kWh/h)	1.161	203	1.294	124	0.121 *	1.202	72	0.044
	(0.69)		(0.73)		(1.52)	(0.62)		(0.542)
Peak usage (kWh/h)	1.432	203	1.551	124	0.107	1.432	72	-0.038
	(1.11)		(1.03)		(0.89)	(0.96)		(-0.298)
TOU Rate (1=yes)	0.182	203	0.202	124	0.016	0.181	72	0.056
	(0.39)		(0.40)		(0.37)	(0.39)		(1.153)
Home ownership (1=yes)	0.774	199	0.821	123	0.030	0.855	69	0.005
	(0.42)		(0.39)		(0.64)	(0.36)		(0.091)
Annual income ($1000)	72.00	199	75.00	123	0.002	76.00	69	-0.001
	(29.00)		(28.00)		(0.69)	(28.00)		(-0.181)
Home size (1000 square feet)	1.541	185	1.908	114	0.351 **	1.611	66	-0.078
	(1.10)		(1.84)		(2.10)	(1.16)		(-0.550)
Age of home (years)	52.221	154	56.574	94	5.195	53.375	56	-0.184
	(30.43)		(31.02)		(1.31)	(28.59)		(-0.044)

*Notes: Means are reported by treatment group, with standard deviations in parantheses below. "Difference" displays the difference in means between each treatment group and control, with t-stats reported in parantheses below. *, **, *** denote significant at the 0.10, 0.05, and 0.01 level.*

Table 3: Group Assignment Balance on Observables, Initial and Compliers

	Initial Group		Compliers	
	Price	Price + IHD	Price	Price + IHD
Mean Off Peak kWh	0.062	0.006	0.030	0.053
	(0.041)	(0.043)	(0.029)	(0.073)
TOU Rate (1=yes)	-0.021	0.073	-0.018	-0.261**
	(0.074)	(0.071)	(0.053)	(0.110)
F-Statistic	1.198	0.673	0.568	2.825
P-Value	0.134	0.882	0.298	0.466
N	337	307	130	100

*Notes: Results denoted "Initial Group" are from a linear probability model regressing observables on the treatment group indicator. Results denoted "Compliers" are from a LPM regressing observables on a compliance indicator. P-Value corresponds to probability that coefficient equal zero. Control group used as control in each specification. Standard errors in parantheses. *,**,*** denote significant at the 0.10, 0.05, and 0.01 level. level.*

Table 4: Survey Compliance by Treatment and Control Group

	Pre-Survey		Post-Survey	
	Price	Price+IHD	Price	Price + IHD
Mean Off-Peak kWh	0.003	-0.025	-0.113*	-0.052
	(0.012)	(0.028)	(0.060)	(0.072)
Rate RT (1=yes)	0.007	-0.003	-0.004	0.025
	(0.022)	(0.043)	(0.110)	(0.108)
F-Statistic	0.158	0.466	2.409	0.256
P-Value	0.794	0.377	0.060	0.476
N	130	100	130	100

*Notes: Results are from a linear probability model regressing observables on pre- and post-survey compliance indicator. The sample in each specification is the initial treatment group. P-Value corresponds to probability that coefficient equals zero. Standard errors in parantheses. *,**,*** denote significant at the 0.10, 0.05, and 0.01 level. level.*

Table 5: Mean kWh Differences (wrt Control) by Treatment Group

		Mean kWh During Events			Difference in Mean kWh wrt Control	
Event Type	Variable	Control	Price	Price + IHD	Price	Price + IHD
Sample: Unbalanced Panel						
DA	Mean	1.65	1.59	1.35	-0.06	-0.30 *
	Std Dev	(1.51)	(1.25)	(1.22)		
	Obs	207	130	100		
TM	Mean	2.07	1.99	1.79	-0.07	-0.28
	Std Dev	(1.77)	(1.54)	(1.42)		
	Obs	186	128	87		
Sample: Balanced Panel						
DA	Mean	1.79	1.67	1.54	-0.13	-0.25 *
	Std Dev	(1.56)	(1.13)	(1.24)		
	Obs	172	90	77		
TM	Mean	2.17	2.17	1.92	0.00	-0.25
	Std Dev	(1.79)	(1.39)	(1.44)		
	Obs	172	90	77		

*Notes: This table presents raw household-level means of hourly kWh during each 15-minute interval, and their difference (for treatment groups with respect to control) during price event periods of each type (DA and TM). The "Unbalanced Panel" and "Balanced Panel" samples are comprised of households assigned to a treatment for which we observe usage data for "at least one pricing event" and "all pricing events", respectively. Standard deviations in parentheses. * Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level.*

Table 6: Intent-to-Treat Effects (Unbalanced Panel)

Event Type:	Pooled	Pooled	Pooled	Pooled	Day Ahead (DA)	30min (TM)
Column:	(1)	(2)	(3)	(4)	(5)	(6)
Price Only	-0.031	-0.054	-0.027	-0.038	-0.071*	0.006
	(0.036)	(0.036)	(0.036)	(0.036)	(0.042)	(0.044)
Price + IHD	-0.116**	-0.137***	-0.123***	-0.137***	-0.171***	-0.084
	(0.048)	(0.048)	(0.047)	(0.046)	(0.051)	(0.057)
Prob(P = P+I)	0.096*	0.098*	0.051*	0.044**	0.066*	0.130
Hour-by-day FEs	N	Y	N	Y	Y	Y
HH FEs	N	N	Y	Y	Y	Y
Number of Events	6	6	6	6	3	3
Number of HHs	437	437	437	437	437	401
R-Square	0.00	0.05	0.54	0.58	0.58	0.58

*Notes: Results are reported from an OLS regression where the dependent variable is ln(kwh) in 15-minute intervals. The sample is comprised of households assigned to a treatment for which we observe usage data for AT LEAST ONE pricing event. All specifications include a treatment group indicator and an event window indicator (except where subsumed by time or household fixed effects). In columns 1-4 the treatment window indicator is set equal to 1 if a DA or TM event is occurring. Column 2 includes hour-by-day fixed effects; column 3 includes household fixed effects and column 4 includes both. In column 5 , the treatment window is set equal to 1 only for DA events and in column 6 the treatment window is set equal to 1 only for TM events. Standard errors in parentheses are clustered at the household level. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 7: Intent-to-Treat Effects (Balanced Panel)

Event Type:	All	All	All	All	Day Ahead (DA)	30min (TM)
Column:	(1)	(2)	(3)	(4)	(5)	(6)
Price Only	-0.009	-0.041	0.001	-0.025	-0.064	0.026
	(0.041)	(0.041)	(0.041)	(0.040)	(0.047)	(0.050)
Price + IHD	-0.109**	-0.132***	-0.105**	-0.124**	-0.153***	-0.079
	(0.050)	(0.050)	(0.049)	(0.048)	(0.052)	(0.058)
Prob(P = P+I)	0.070*	0.098*	0.051*	0.065*	0.126	0.100*
Hour-by-day FEs	N	Y	N	Y	Y	Y
HH FEs	N	N	Y	Y	Y	Y
Number of Events	6	6	6	6	3	3
Number of HHs	339	339	339	339	339	339
R-Square	0.00	0.06	0.51	0.56	0.56	0.56

*Notes: Results are reported from an OLS regression where the dependent variable is ln(kwh) in 15-minute intervals. The sample is comprised of households assigned to a treatment for which we observe usage data for ALL pricing events. All specifications include a treatment group indicator and an event window indicator (except where subsumed by time or household fixed effects). In columns 1-4 the treatment window indicator is set equal to 1 if a DA or TM event is occurring. Column 2 includes hour-by-day fixed effects; column 3 includes household fixed effects and column 4 includes both. In column 5 , the treatment window is set equal to 1 only for DA events and in column 6 the treatment window is set equal to 1 only for TM events. Standard errors in parentheses are clustered at the household level. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 8: Treatment Effect on Treated Households (Unbalanced Panel)

Event Type:	Pooled	Pooled	Pooled	Pooled	Day Ahead (DA)	30min (TM)
Column:	(1)	(2)	(3)	(4)	(5)	(6)
Price Only	-0.032	-0.056	-0.028	-0.040	-0.074*	0.007
	(0.037)	(0.037)	(0.037)	(0.037)	(0.044)	(0.046)
Price + IHD	-0.143**	-0.170***	-0.153***	-0.170***	-0.217***	-0.100
	(0.058)	(0.058)	(0.057)	(0.057)	(0.064)	(0.067)
Prob(P = P+I)	0.061*	0.052*	0.030**	0.023**	0.025**	0.115
Hour-by-day FEs	N	Y	N	Y	Y	Y
HH FEs	N	N	Y	Y	Y	Y
Number of Events	6	6	6	6	3	3
Number of HHs	437	437	437	437	437	401
R-Square	0.00	0.05	0.54	0.58	0.58	0.58

*Notes: Results are reported from a 2SLS regression where the dependent variable is ln(kwh) in 15-minute intervals. Initial treatment assignment is used as an instrument for receipt of treatment. The sample is comprised of households assigned to a treatment for which we observe usage data for AT LEAST ONE pricing event. All specifications include a treatment group indicator and an event window indicator (except where subsumed by time or household fixed effects). In columns 1-4 the treatment window indicator is set equal to 1 if a DA or TM event is occurring. Column 2 includes hour-by-day fixed effects; column 3 includes household fixed effects and column 4 includes both. In column 5 , the treatment window is set equal to 1 only for DA events and in column 6 the treatment window is set equal to 1 only for TM events. Standard errors in parentheses are clustered at the household level. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 9: Treatment Effect on Treated Households (Balanced Panel)

Event Type:	Pooled	Pooled	Pooled	Pooled	Day Ahead (DA)	30min (TM)
Column:	(1)	(2)	(3)	(4)	(5)	(6)
Price Only	-0.009	-0.042	0.001	-0.025	-0.066	0.026
	(0.041)	(0.042)	(0.041)	(0.041)	(0.048)	(0.051)
Price + IHD	-0.129**	-0.156***	-0.124**	-0.147***	-0.182***	-0.094
	(0.058)	(0.058)	(0.057)	(0.056)	(0.061)	(0.068)
Prob(P = P+I)	0.051*	0.062*	0.037**	0.042**	0.072*	0.091*
Hour-by-day FEs	N	Y	N	Y	Y	Y
HH FEs	N	N	Y	Y	Y	Y
Number of Events	6	6	6	6	3	3
Number of HHs	339	339	339	339	339	339
R-Square	0.00	0.06	0.51	0.56	0.56	0.56

*Notes: Results are reported from a 2SLS regression where the dependent variable is ln(kwh) in 15-minute intervals. Initial treatment assignment is used as an instrument for receipt of treatment. The sample is comprised of households assigned to a treatment for which we observe usage data for ALL pricing events. All specifications include a treatment group indicator and an event window indicator (except where subsumed by time or household fixed effects). In columns 1-4 the treatment window indicator is set equal to 1 if a DA or TM event is occurring. Column 2 includes hour-by-day fixed effects; column 3 includes household fixed effects and column 4 includes both. In column 5 , the treatment window is set equal to 1 only for DA events and in column 6 the treatment window is set equal to 1 only for TM events. Standard errors in parentheses are clustered at the household level. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 10: Notification of Treatment Events

	Price Only				Price + IHD			
Event Date	Confirmed	Intermediate	Not Confirmed	Total	Confirmed	Intermediate	Not Confirmed	Total
07/21/11	45	29	27	101	34	29	17	80
	45%	29%	27%		43%	36%	21%	
07/22/11	37	36	30	103	28	37	15	80
	36%	35%	29%		35%	46%	19%	
08/04/11	42	58	25	125	26	44	15	85
	34%	46%	20%		31%	52%	18%	
08/10/11	49	45	30	124	23	45	18	86
	40%	36%	24%		27%	52%	21%	
08/17/11	45	52	30	127	21	49	17	87
	35%	41%	24%		24%	56%	20%	
08/26/11	49	48	32	129	24	46	30	100
	38%	37%	25%		24%	46%	30%	

Notes: A household is classified as "Confirmed" if it received text message confirmations, or if it actively confirmed the receipt of the event notification by phone or email. "Intermediate" households were recorded as having received a phone call or voicemail, but did not press the "confirm" button at the conclusion of the automated notification call/voicemail. All remaining households including households who did not confirm receipt of an email are classified as "Not Confirmed".

Table 11: Notification Confirmation and Treatment Effects

	Intermediate and Confirmed Separate			Confirmed = Confirmed + Intermediate		
Event Type:	All events	DA events	TM events	All events	DA events	TM events
Column:	(1)	(2)	(3)	(4)	(5)	(6)
Price*1[Not Confirmed]	-0.007	-0.043	0.038	-0.007	-0.043	0.038
	(0.048)	(0.066)	(0.057)	(0.048)	(0.066)	(0.057)
Price+IHD*1[Not Confirmed]	-0.050	-0.104	0.037	-0.050	-0.104	0.037
	(0.080)	(0.087)	(0.110)	(0.080)	(0.087)	(0.110)
Price*1[Intermediate]	-0.010	-0.040	0.030			
	(0.049)	(0.059)	(0.066)			
Price+IHD*1[Intermediate]	-0.163***	-0.164***	-0.148**			
	(0.053)	(0.057)	(0.074)			
Price*1[Confirmed]	-0.086*	-0.116*	-0.041	-0.049	-0.080*	-0.005
	(0.052)	(0.061)	(0.068)	-0.040	(0.046)	(0.051)
Price+IHD*1[Confirmed]	-0.162*	-0.231**	-0.053	-0.162***	-0.192***	-0.113*
	(0.083)	(0.100)	(0.082)	(0.052)	(0.057)	(0.062)
P-Value (PIHD*NC = P*NC)	0.628	0.558	0.990	0.628	0.557	0.991
P-Value (PIHD*C = P*C)	0.414	0.302	0.900	0.047**	0.073*	0.120
HH FEs	Yes	Yes	Yes	Yes	Yes	Yes
Hour-by-day FEs	Yes	Yes	Yes	Yes	Yes	Yes
Number of hhs	437	437	401	437	437	401
R-Square	0.583	0.583	0.583	0.583	0.583	0.583

*Notes: Results are reported from an OLS regression where the dependent variable is ln(kwh) and the treatment indicator is interacted wtih notification confirmation category. The sample is comprised of households assigned to a treatment for which we observe usage data for AT LEAST ONE pricing event. In columns 1-3, "Intermediate" and "Confirmed" are estimated as separate notification confirmation categories; in columns 4-6 "Intermediate" is re-classified as "Confirmed". P-Value reports probability of equal effects across groups. Standard errors in parentheses are clustered by household. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 12: Home During Daytime

	% of HHs: Initially Assigned	% of HHs: Compliers Only	Intention to Treat			Treatment on the Treated		
			All events	DA events	TM events	All events	DA events	TM events
Price*1[Home during day]	51%	52%	-0.036	-0.056	-0.009	-0.032	-0.052	-0.005
			(0.043)	(0.052)	(0.051)	(0.044)	(0.053)	(0.052)
Price*1[Home after school]	22%	20%	-0.042	-0.095	0.028	-0.041	-0.1	0.035
			(0.056)	(0.060)	(0.083)	(0.063)	(0.067)	(0.090)
Price*1[Home after work]	27%	27%	-0.031	-0.068	0.017	-0.027	-0.065	0.022
			(0.067)	(0.080)	(0.080)	(0.069)	(0.082)	(0.082)
Price*1[Missing]	1%	0%	-0.426***	-0.957***	0.636***			
			(0.027)	(0.034)	(0.038)			
IHD+price*1[Home during day]	41%	44%	-0.142**	-0.179**	-0.086	-0.173**	-0.221**	-0.103
			(0.072)	(0.079)	(0.083)	(0.088)	(0.097)	(0.103)
IHD+price*1[Home after school]	21%	22%	-0.124*	-0.13	-0.109	-0.127*	-0.14	-0.105
			(0.071)	(0.091)	(0.082)	(0.076)	(0.102)	(0.082)
IHD+price*1[Home after work]	35%	29%	-0.123*	-0.174**	-0.045	-0.164*	-0.244**	-0.054
			(0.067)	(0.072)	(0.086)	(0.094)	(0.104)	(0.113)
IHD+price*1[Missing]	3%	4%	-0.275	-0.28	-0.25	-0.420***	-0.950***	0.640***
			(0.300)	(0.235)	(0.365)	(0.027)	(0.033)	(0.037)
P-Value (PIHD*HDD = P*HDD)			0.174	0.154	0.386	0.120	0.092*	0.350
HH FEs			Yes	Yes	Yes	Yes	Yes	Yes
Hour-by-day FEs			Yes	Yes	Yes	Yes	Yes	Yes
Number of obs			230	230	215	230	230	215
R-Square			0.571	0.571	0.571	0.571	0.571	0.571

*Notes: Treatment effects by pre-existing daily schedule. P-Value reports probability of equal treatment effects across groups, conditional on being "generally home during the day". Standard errors clustered at the household in parantheses. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 13: Central Air Conditioning

	% of HHs	All events	DA events	TM events
Price*1[No Central AC]	55%	-0.090**	-0.085*	-0.089*
		(0.042)	(0.048)	(0.053)
Price*1[Yes Central AC]	42%	0.035	-0.051	0.144**
		(0.049)	(0.064)	(0.057)
Price*1[Missing]	4%	-0.078	-0.071	-0.081
		(0.143)	(0.132)	(0.190)
IHD+P*1[No Central AC]	61%	-0.184***	-0.226***	-0.117*
		(0.054)	(0.059)	(0.071)
IHD+P*1[Yes Central AC]	35%	-0.071	-0.083	-0.052
		(0.073)	(0.082)	(0.079)
IHD+P*1[Missing]	4%	-0.048	-0.17	0.123
		(0.445)	(0.420)	(0.297)
P-Value (P*Yes CAC = PIHD*Yes CAC)		0.191	0.744	0.0259**
P-Value (P*No CAC = PIHD*No CAC)		0.124	0.032**	0.722
HH FEs		Yes	Yes	Yes
Hour-by-day FEs		Yes	Yes	Yes
Number of hhs		230	230	215
R-Square		0.571	0.571	0.571

*Notes: Treatment effects by presence of central air conditioning. P-Value reports probability of equal treatment effects across groups, conditional on the indicated observable. Standard errors clustered by household in parantheses. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 14: Frequency of IHD Interaction

	% of HHs	All events	DA events	TM events
Price+IHD*1[0/None]	4%	-0.366***	-0.585***	-0.086
		(0.111)	(0.154)	(0.260)
Price+IHD*1[1-2 times]	10%	-0.008	0.003	-0.022
		(0.082)	(0.084)	(0.100)
Price+IHD*1[3-5 times]	8%	0.065	0.042	0.089
		(0.070)	(0.066)	(0.079)
Price+IHD*1[More than 5 times]	40%	-0.241***	-0.256***	-0.208***
		(0.065)	(0.072)	(0.076)
Price+IHD*1[Missing]	38%	-0.007	-0.031	0.027
		(0.074)	(0.077)	(0.095)
P-Value (PIHD*>5 = PIHD*1-2)		0.022**	0.015**	0.128
P-Value (PIHD*>5 = PIHD*3-5)		0.001***	0.001***	0.005***
HH FEs		Yes	Yes	Yes
Hour-by-day FEs		Yes	Yes	Yes
Number of obs		100	100	87
R-Square		0.571	0.571	0.571

*Notes: Treatment effect and survey-reported initial weekly frequency of IHD interaction. P-Value reports probability of equal treatment effects across frequency of experience with IHDs. Standard errors clustered by household in parantheses. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 15: Load Shifting: Anticipation and Spillovers

	DA events	TM events
Price-Only: 2hrs Pre-Event	0.003	0.053
	(0.039)	(0.038)
Price-Only: 2hrs Post-Event	-0.043	-0.051
	(0.047)	(0.046)
Price+IHD: 2hrs Pre-Event	-0.095**	-0.024
	(0.042)	(0.045)
Price+IHD: 2hrs Post-Event	-0.103*	-0.027
	(0.055)	(0.057)
HH FEs	Yes	Yes
Hour-by-day FEs	Yes	Yes
Number of hhs	437	401
R-Square	0.568	0.568

*Notes: The specification is the baseline ITT with additional regressor indicator variables for 2-hrs pre- and post-treatment event. The sample includes households that were present for at least one treatment event (what we are calling the unbalanced panel). The specification includes treatment indicators, coefficients for which are not reported. Standard errors clustered by household in parantheses. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 16: Habit Formation

	Price	Price + IHD
12-1pm Calendar Day Trend	-0.0019	-0.0020
	(0.0015)	(0.0015)
1-2pm Calendar Day Trend	-0.0020	-0.0020
	(0.0015)	(0.0015)
2-3pm Calendar Day Trend	-0.0020	-0.0018
	(0.0015)	(0.0014)
3-4pm Calendar Day Trend	-0.0018	-0.0020
	(0.0014)	(0.0014)
4-5pm Calendar Day Trend	-0.0020	-0.0025*
	(0.0014)	(0.0013)
5-6pm Calendar Day Trend	-0.0025*	-0.0023
	(0.0013)	(0.0014)
6-7pm Calendar Day Trend	-0.0023	-0.0026**
	(0.0014)	(0.0013)
7-8pm Calendar Day Trend	-0.0026**	-0.0028**
	(0.0013)	(0.0014)
HH FEs	Yes	
Hour-by-day FEs	Yes	
Number of hhs	339	
R-Square	0.560	

*Notes: Results from a single regression specification which interacts a calendar day time trend for each peak hour with initial treatment assignment. The sample is restricted to all non-pricing event weekdays in July and August, and includes only households that were present for all treatment events (what we are calling the balanced panel). Standard errors clustered by household in parantheses. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*

Table 17: Annual Value of Conservation During Price Events

	DA, $0.50/kWh			TM, $1.25/kWh		
	Price	Price+IHD	Difference	Price	Price+IHD	Difference
20hrs	$1.17	$3.24	$2.07	$0.08	$4.01	$3.93
40hrs	$2.33	$6.48	$4.15	$0.16	$8.02	$7.87
60hrs	$3.50	$9.72	$6.22	$0.24	$12.04	$11.80

Estimated from average peak kWh usage and estimated treatment effects by event type.

Figure A-1: In-Home Display (1)

Figure A-2: In-Home Display (2)

Table A-1: ITT and ToT Specifications Weighted to Reflect US Income Distribution

Estimator:	ITT			ToT		
Event Type:	All Events	Day Ahead (DA)	30min (TM)	All Events	Day Ahead (DA)	30min (TM)
Price Only	-0.013	-0.031	0.012	-0.013	-0.032	0.012
	(0.048)	(0.058)	(0.061)	(0.050)	(0.061)	(0.062)
Price + IHD	-0.133**	-0.189**	-0.047	-0.178**	-0.267**	-0.059
	(0.061)	(0.083)	(0.109)	(0.086)	(0.116)	(0.137)
Prob(P = P+I)	0.049**	0.070*	0.588	0.042**	0.041**	0.597
Hour-by-day FEs	Y	Y	Y	Y	Y	Y
HH FEs	Y	Y	Y	Y	Y	Y
Number of Events	6	3	3	6	3	3
Number of HHs	429	429	393	429	429	393
R-Square	0.57	0.57	0.57	0.57	0.57	0.57

*Notes: Reported results are from the ITT and ToT specifications, weighted to reflect the national income distribution. The sample is the unbalaced panel. Standard errors in parentheses are clustered at the household level. *,**,*** indicates significance at 0.10, 0.05, and 0.01.*