Federal Reserve Bank of Minneapolis
Research Department Staff Report 360

March 2005

Intellectual Property and Market Size*

Michele Boldrin

Federal Reserve Bank of Minneapolis
and University of Minnesota

David K. Levine

University of California, Los Angeles
and Federal Reserve Bank of Minneapolis

ABSTRACT

Intellectual property protection involves a trade-off between the undesirability of monopoly and the desirable encouragement of creation and innovation. As the scale of the market increases, due either to economic and population growth or to the expansion of trade through treaties such as the World Trade Organization, this trade-off changes. We show that, generally speaking, the socially optimal amount of protection decreases as the scale of the market increases. We also provide simple empirical estimates of how much it should decrease.

*This paper is based upon work supported by the National Science Foundation under Grants SES 01-14147 and 03-14713. Boldrin also acknowledges research support from the Spanish BEC2002-04294-C02-01. Our ideas benefited from comments at the Theory Workshop at Columbia University, the brown bag workshop at the Federal Reserve Bank of Dallas, the faculty seminar at the Chinese University of Hong Kong, the Theory Workshop at ASU, the CSIP seminar at the Federal Reserve Bank of San Francisco, and from discussions with Kyle Bagwell, Chad Jones, Pete Klenow, and Daniel Wilson. Thanks are also due to Hengjie Ai's research assistance in collecting and analyzing data. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

The U.S. Constitution gives Congress the power "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." This recognizes the two basic economic features of intellectual property protection: On the one hand, exclusive rights create monopoly power and so should be limited in time. On the other, monopoly power provides an incentive for creation and innovation. For practical reasons the same time limit applies across a wide variety of creations and innovations: In U.S. law, copyright is life of author plus 70 years for individual works and 95 years for works for hire. Design patents are 20 years, and ornamentation patents are 14 years. Since the private profitability of creating and innovating varies widely, this means that for any fixed time limit many ideas will earn profits above the level needed to recoup the cost of innovation. In a larger market profits will be greater, and inframarginal ideas will earn additional economically unnecessary rents. Hence, as the market expands, it becomes possible to reduce the length of term without reducing the production of new ideas. But, as the market expands, some ideas that were not profitable to produce will become so, and reducing the length of term will discourage these marginal entrepreneurs. Which of these two competing forces should matter more for good policy?

In this paper we look at the general equilibrium interaction that determines how optimal protection varies with the scale of the market. Profitability of an innovation depends upon three factors: the initial cost of discovery, the elasticity of demand, and, finally, the size of the market. All these elements vary widely and unsystematically across innovations. We concentrate on market size for three reasons. Contrary to the other two factors, market size is straightforward to measure. Secondly, growth in per capita income and the expansion of international trade have increased market size by two orders of magnitude since U.S. patent and copyright legislation was introduced. Finally, the ongoing process of trade expansion has put the international harmonization of intellectual property rights at center stage, through the World Trade Organization's agreement on trade-related aspects of intellectual property rights (WTO-TRIPS).

Our basic result is an intuitive one. Optimal policy involves a tradeoff between increasing the monopolistic distortion on inframarginal ideas and increasing the number of marginal ideas. As the scale of the market increases, it will generally be desirable to give up some of the additional marginal ideas in exchange for reduction of monopoly across the broad variety of inframarginal ideas that will be produced anyway, and so the optimal policy should reduce the length of protection as the scale of the market increases.

We make this point in the context of a simple model, intentionally designed to render intellectual property protection socially beneficial. Ideas are created subject to an indivisibility, or fixed cost. There are many possible

ideas, and to model the fact that each should have downward sloping demand, we adopt the Dixit-Stiglitz model of preferences. The *private return* on an idea is the ratio of expected monopoly revenue to its cost of creation. We consider first the case in which the private return has a neutral effect on the relationship between the private and social benefit of an idea. We show that the complex heterogeneous mass of ideas can be analyzed by examining the total monopoly revenue from all ideas with a private return above a threshold level. Using this tool, we show that when the market is sufficiently small, it may be optimal to provide an unlimited monopoly, but when the market is large enough, a time limit should always be imposed, and this limit should strictly decrease as the size of the market grows.

This model is related to a series of papers by Grossman and Helpman [1991, 1994, 1995] studying innovation in a Dixit-Stiglitz framework. It is most closely related, however, to Grossman and Lai [2002, 2004]. Their approach differs from ours in two respects. First, where we use a static analysis, they embed the static model in a dynamic setting by treating costs and profits as time-flows. While this approach does not answer all dynamic issues, such as the depletion of existing ideas and ideas that use other ideas as inputs, it is a valid dynamic interpretation of the model we use. Since Grossman and Lai have already provided this interpretation, we do not do so here. Second, their model uses a production function approach to the creation of new ideas. That is, ideas are of homogeneous quality and are produced using a constant returns technology with human capital and labor as inputs. Although this approach to the production of ideas is less versatile than our disaggregated model of heterogeneous ideas, under the assumption of symmetry it is possible to translate production functions into equivalent total monopoly revenue functions, so in this case their model has the same reduced form as ours, and we show how to make the connection.

The most significant difference between our results and those of Grossman and Lai are that they focus on the case in which the production function is Cobb-Douglas. Both they and we show that this means that optimal protection does not change with the size of the market. If, though, the total monopoly revenue function has increasing elasticity, then optimal protection locally decreases and vice versa. The Cobb-Douglas case is then the boundary between these two general cases. Although we also show, as noted above, that when market size is large enough, optimal protection must always decrease—even in the Cobb-Douglas case—the question arises: Are we currently in a region of the total monopoly revenue function where optimal protection might be constant or increasing? We give both theoretical reasoning and empirical evidence that this is not the case.

Grossman and Lai focus also on "harmonization" and North-South trade. Because they provide a detailed treatment, especially of issues involving who benefits from harmonization, we examine only the baseline case involving multiple countries. Our goal is to show how—contrary to their finding—in the empirically relevant case of increasing elasticity of total monopoly

revenue, the North should reduce protection as a result of harmonization. In the case of two countries of equal size, because some of the benefits of a higher time limit are received by the other country there is a tendency to set protection too low, and there is a "harmonization" argument to be made for international treaties raising the time limit. However, this argument applies only to countries of equal size. When the countries, two or more, are of unequal size, the smaller country tends to set low limits and free ride off the large country—but the large country tends to set limits that are too high because it does not account for the social benefit of innovation to the smaller country. An implication of this result is that the process of trade expansion should be accompanied by a parallel process of intellectual property reduction. In this case "harmonization" does not mean setting limits equal to or higher than those in the larger and more protected country, but rather adjusting the time limits to lie in between the larger protection of the larger country and the smaller protection of the smaller country.

The general equilibrium approach emphasizes the connection between broader features of the economy and intellectual property. We illustrate this through various comparative static results. Increasing the scale of the market increases the demand for the specialized labor needed to create ideas. For given intellectual property (IP) protection, this drives up the wages of that kind of labor. These increased economic rents do not serve a useful economic purpose, because they do not increase the number of ideas that are produced; their only effect is to redistribute income from the rest of the economy to the subset of IP protected workers. We also consider a number of extensions of the basic model. The most important is to relax the assumption that quality is neutral in the relation between private and social values of an idea. As either the scale of the market or the term of intellectual property protection increases, ideas that are more marginal from a private point of view are produced. If the social value of these ideas declines even faster, then the argument for decreasing the length of protection with market scale is strengthened. If ideas that are more marginal from a private perspective are more beneficial from a social perspective, then a system of exclusive rights is a poor method of encouraging the production of valuable ideas: it leads to the production of the least, rather than the most, socially useful ideas. When legislation is in place that allocates monopoly power to the producers of certain goods and not of others (patent and copyright protection varies widely across economic sectors), a natural consequence is the emergence of socially wasteful lobbying and rent-seeking activities. In our framework rent-seeking would consist of payments going from producers to legislators to increase the length of intellectual property protection. The basic result is that the entrepreneurs who have access to the less marginal projects are more willing to pay for the legal monopoly to be continued or extended. Hence, in the presence of rent-seeking behavior, intellectual property protection leads to an equilibrium in which its length is determined, at the margin, by those innovators who need it the least.

One may wonder why the relation between the size of the market and the optimal level of IP protection (that is, the optimal degree of monopoly power) is not emphasized in the "new" growth literature, which has at its core a model of adoption of new commodities similar to the one used here. We see two reasons: First, the question is never asked because that literature assumes that monopoly power is good and necessary for innovation. Second, because the impact of market size on the optimal length of IP protection is buried behind the so-called size effect, according to which increasing the size of the market speeds up the growth process due to increasing returns. So, for example, in the Romer [1990] paper on endogenous technological change, one reads that (p. S95) "... these models have an underlying form of increasing returns in research. As a result, an increase in a scale variable induces an increase in the rate of growth."

2. The Model

Ideas are indexed by their characteristics ω, which measure the cost and utility of an idea and lie in Ω, a compact subset[1] of $\Re^n$. To be invented, each idea requires a minimum amount $h(\omega) \geq 0$ of the only primary input, labor, where $h(\omega)$ is a measurable function. We refer to $h(\omega)$ as the indivisibility, minimum size, or fixed cost for producing a new idea. The "number" of ideas with given characteristics in an economy of unit size is a positive measure $\eta(\omega)$. We will later focus on the case where $\eta(\omega)$ is a probability distribution, and innovators find their individual ideas by drawing from this underlying distribution, but this interpretation is not essential. Allowing numerous ideas with the same characteristics is useful because it makes it easy to think about the possibility that doubling the size of an economy might double the number of ideas of given cost and utility.

There is a continuum population of size λ of agents; with λ we measure the scale of the economy. The number of available ideas may depend on the size of the economy, so the total number of ideas with characteristics ω available in an economy of size λ is $g(\lambda)\eta(\omega)$. To capture the principle that in a larger population more ideas of a given quality are available, $g(\lambda)$ is assumed nondecreasing in λ; we may assume without loss of generality that $g(1) = 1$. In the case in which $\eta(\omega)$ is a probability distribution from which innovators draw their ideas without replacement, twice as many innovators means twice as many ideas with given characteristics, so $g(\lambda) = \lambda$. Neither that the number of ideas increases with size at different rates for different characteristics nor that the indivisibility varies with the size of the economy is a possibility considered here.

We assume initially that once an idea is created, it may be reproduced at no cost and without limit. If the input of labor $y(\omega)$ used in producing ω is below the threshold, that is, $y(\omega) < h(\omega)$, no prototype will emerge and no

[1] Actually, any topological measure space will do.

consumption is possible. If $y(\omega) \geq h(\omega)$, then consumption is $x(\omega) \geq 0$. It is convenient also to measure consumption per capita as $z(\omega) = x(\omega)/\lambda$.

The utility of ideas is uncertain at the time the invention decision is made.[2] In our model, this means that it is the expected return on an idea *ex ante,* and not the *ex post* determination of whether the idea turns out to be a good one or a bad one, that matters for the decision to invent. For concreteness, we may imagine that z units of an idea with characteristics ω has utility to a representative consumer of $u(z, \omega) \geq 0$ with probability $p(\omega)$, while with probability $1 - p(\omega)$ the idea has no utility at all.[3] Normalize $u(0, \omega) = 0$, and assume that $p(\omega), u(z, \omega)$ are continuous in ω with the latter also continuous and nondecreasing in z and, at least up to a limit $Z(\omega)$, smooth and strictly increasing.

Set $v(z, \omega) = p(\omega)u(z, \omega)$ to be expected utility. We assume $\lim_{z\to\infty} v(z, \omega) = v^C(\omega) < \infty$. Since $v(z, \omega)$ is bounded, $zv_z(z, \omega) \to 0$ as $z \to \infty$; that is, per capita revenue falls to zero as per capita consumption grows without bound. We also assume that $zv_z(z, \omega)$ has a unique maximum at $z^M(\omega)$.

The utility of a representative individual has a Dixit-Stiglitz form over goods of different characteristics. Apart from consumption of idea-goods, consumers receive utility from time ℓ spent on activities that take place outside of the idea sector. If L is the individual endowment of time, $0 \leq \ell \leq L$. Since $g(\lambda)\eta(\omega)$ of type ω ideas are potentially available, individual utility is

$$\int v(z(\omega), \omega)g(\lambda)\eta(d\omega) + \ell.$$

Note that the marginal utility of time outside the idea sector is normalized to one. We will consider later the possibility that this marginal utility changes with the scale of the economy; for example, increases in per capita GDP may increase the productivity of labor outside the idea sector. The social feasibility constraint is that the amount of time spent outside the idea sector equals the amount left over after the production of ideas

$$\lambda(L - \ell) = \int y(\omega)g(\lambda)\eta(d\omega).$$

Profit maximization and efficiency require $y(\omega) = h(\omega)$ for all ideas for which $x(\omega) > 0$, and $y(\omega) = 0$, otherwise. It is also obvious that no good would be produced in this economy absent patent protection; we look later at the less extreme case in which a capacity constraint exists and making copies involves a positive marginal cost.

---

[2]The cost might be uncertain as well, and in a dynamic setting it may also be possible to develop an idea gradually, choosing to stop if the idea turns out to be a poor one. We abstract from these difficulties.

[3]Other functional forms can be accommodated; more general distributional assumptions add notational complexity, with little substantive implications for our model or results.

Patent Equilibrium. Our notion of equilibrium is that of a *patent equilibrium* in which there is a fixed common length of patent protection for all ideas. This means that, in terms of present value of the flow of consumption, a fraction $0 \leq \phi \leq 1$ occurs under monopoly, and a fraction $(1-\phi)$ occurs under competition; hence, ϕ is the level or the extent of protection. Potentially, many individuals can invent any particular idea; certainly the number of individuals who have historically had truly unique ideas is minuscule. We do not model the "patent race" by which patent is awarded, and simply assume that, for each of the $\eta(\omega)$ ideas with characteristics ω, a particular individual is awarded a "patent." While the patent lasts, the inventor is a monopolist, and our economy is similar to the traditional Dixit-Stiglitz "monopolistic competition" economy. Once a patent expires, anyone who wishes to do so may make copies of ideas that had been previously introduced under the patent regime. Once competition sets in, output and consumption jump to infinity while prices and revenue fall to zero.[4] A type of good is produced if, given the patent length ϕ, the prospective monopolist finds it profitable to overcome the indivisibility. This notion of equilibrium is closely connected to that of Hart [1979] and Makowski [1980], and it has been used, for example, by Acemoglu and Zilibotti [1996] in a related context.

The market for innovation is equilibrated through the wage rate of labor w.. The higher is w, the costlier it is to produce new ideas, and fewer of them will therefore be produced. If the amount of labor used in the production of ideas is strictly less than the total endowment λL, wages $w = 1$. Otherwise, w must be chosen to reduce demand for labor to the point where the amount of leisure is 0.

A monopolist who holds a patent for a good with characteristics ω and sells z units of output to each of the λ consumers receives revenue $\lambda z(\omega) v_z (z(\omega), \omega)$, which is assumed to have a unique maximum at $z^M(\omega)$, and pays the cost $wh(\omega)$. For a commodity with characteristics ω, $\rho(\omega) = z^M(\omega) v_z(z^M(\omega), \omega)/h(\omega)$ expresses the ratio of (per capita) private value to the innovation cost. In fact, $\rho(\omega)/w$ represents one plus the rate of return on investment which would accrue to the inventor of commodity ω if patents lasted forever and the market size was $\lambda = 1$. We refer to $\rho(\omega)$ as the *private return* for ω. The monopolist receives a fraction ϕ of the private return, times the size λ of the market. Hence, a good is produced if

$$\rho(\omega) \geq w/\phi\lambda \equiv \underline{\rho}.$$

In other words, no ideas with private return lower than $\underline{\rho}$ will be introduced in the patent equilibrium, and all ideas with a $\rho(\omega)$ above $\underline{\rho}$ will be produced. Notice that $\underline{\rho}$ is strictly decreasing in $\phi\lambda$, meaning that as the scale of the market or the extent of protection increases, ideas with a lower private return are introduced. Notice also that, in general, there need not be any

[4] We assume revenue falls to zero as quantity goes to infinity for the sake of simplicity; as pointed out in Boldrin and Levine [1999], it need not be the case, and the inventor will earn positive competitive rents even after the patent expires.

monotone relation between the private return $\rho(\omega)$ of an idea and its social return; hence, ideas of high social return may be introduced only for high values of λ, or even never at all, if their private return $\rho(\omega)$ is particularly low.

Per capita social welfare in a patent equilibrium is derived by integrating utility for those goods that are produced less the cost of producing them

$$\int_{\rho(\omega)\geq\underline{\rho}} [\phi v(z^M(\omega),\omega) + (1-\phi)v^C(\omega) - h(\omega)/\lambda]g(\lambda)\eta(d\omega) + L.$$

We assume that for $\underline{\rho} > 0$

$$L^D = \int_{\rho(\omega)\geq\underline{\rho}}^{\infty} g(\lambda)h(\omega)\eta(d\omega) < \infty,$$

so that the amount of labor required to produce all ideas exceeding any particular private value threshold is finite.

Notice that $\rho(\omega)h(\omega)\eta(\omega)$ is the total revenue of a monopolist investing in goods with characteristics ω in an economy of unit size. For any given cutoff ρ, we can define

$$M(\rho) = \int_{\rho(\omega)\geq\rho}^{\infty} \rho(\omega)h(\omega)\eta(d\omega).$$

Then $M(\rho)$ is the sum of monopoly revenue over all ideas with private value of ρ, or greater. We assume that M is differentiable and define the *elasticity of total monopoly revenue,* with respect to variations in the marginal idea, as $\Upsilon(\rho) \equiv -\rho M'(\rho)/M(\rho) > 0$. We also make the regularity assumption that $\Upsilon(\rho)$ is differentiable.

Let $\nu^M(\omega) \equiv v(z^M(\omega),\omega)/[h(\omega)\rho(\omega)]$ and $\nu^C(\omega) \equiv v^C(\omega)/[h(\omega)\rho(\omega)]$ be the ratio of social value to private return of a commodity of type ω under monopoly and under competition, respectively. To fix ideas, consider the case in which utility has the quadratic form

$$v(\omega,z) = b(\omega)\left(Z(\omega)^2 - [z - Z(\omega)]^2\right)$$

for $z \leq Z(\omega)$ and $v(\omega,z) = b(\omega)Z(\omega)^2$ for $z > Z(\omega)$. Then we have $\nu^M(\omega) = 3/2$ and $\nu^C(\omega) = 2$ independently of characteristics. More generally, we can define the notion of *return neutrality*. If the ratios of social values to private return $\nu^M(\omega)$ and $\nu^C(\omega)$ are both constant, we have *strong return neutrality*. Formally, observe that the measure $h(\omega)\eta(\omega)$ represents, in an economy of unit size, the quantity of labor needed to produce all ideas with characteristics ω.. Consider the measure $h(\omega)\eta(\omega)$, restricted to the σ-subalgebra of the Borel sets of Ω generated by the subsets of Ω on which $\rho(\omega)$ is constant; make the regularity assumption that it can be represented by a continuous density function $\mu(\rho) = \int_{\rho(\omega)=\rho} h(\omega)\eta(d\omega)$. For any function, $f(\omega)$, define a conditional value $\overline{f}(\rho)$ in much the same way as a conditional expectation is defined. Specifically, $\overline{f}(\rho)$ is defined, μ-almost everywhere, by the condition that $\int_B \overline{f}(\rho)\mu(\rho)d\rho = \int_B f(\omega)h(\omega)\eta(d\omega)$ for every set B in the

σ-subalgebra of the Borel sets of Ω on which $\rho(\omega)$ is constant. By *return neutrality* we mean that $\overline{\nu}^M(\rho), \overline{\nu}^C(\rho)$ are constant.[5] Below, we consider first the neutral then the nonneutral case.

3. Return Neutrality

We first examine the case of return neutrality and ask how socially optimal protection $\hat{\phi}(\lambda)$ depends on market size. We find that, if the elasticity of total monopoly revenue is well-behaved near $\rho = 0$, then for large enough λ socially optimal protection must be declining with λ. Further, if the elasticity of total monopoly revenue is increasing with ρ, a condition that, contra Grossman and Lai [2002, 2004], we argue is likely to be the case, then socially optimal protection is in fact decreasing as a function of λ.

Basically, there are two cases. If the elasticity of total monopoly revenue is increasing with ρ and $\hat{\phi}(\lambda) < 1$, we can show from the first-order conditions and implicit function theorem that $\hat{\phi}(\lambda)$ is strictly decreasing. If, instead, the elasticity of total monopoly revenue is decreasing with ρ, then labor demand is growing faster than labor supply, and so the labor constraint must eventually bind. We show that whenever the labor constraint binds, it must be the case that $\hat{\phi}(\lambda)$ is strictly decreasing.

Proposition. *Suppose return neutrality. If for some $\tilde{\rho}$ and $0 < \rho < \tilde{\rho}$, $\Upsilon'(\rho) \neq 0$, then there exists $\overline{\lambda}$ such that $\hat{\phi}(\lambda)$ is unique and strictly decreasing for $\lambda > \overline{\lambda}$. If $\hat{\phi}(\lambda) < 1$, then $\Upsilon'(1/\lambda\widehat{\phi}(\lambda)) > 0$ implies $\hat{\phi}(\lambda)$ is unique and strictly decreasing; $\Upsilon'(\rho) = 0$ in a neighborhood of $1/\lambda\widehat{\phi}(\lambda)$ implies $\hat{\phi}(\lambda)$ is unique and locally constant, and $\Upsilon'(1/\lambda\widehat{\phi}(\lambda)) < 0$ and $\hat{\phi}(\lambda)$ unique[6] implies $\hat{\phi}(\lambda)$ is strictly increasing.*

Proof. Use return neutrality to rewrite social welfare as

$$\int_{\rho' \geq \underline{\rho}} [\phi \overline{\nu}^M \rho' + (1-\phi)\overline{\nu}^C \rho' - 1/\lambda] g(\lambda)\mu(\rho')d\rho' + L.$$

We begin by analyzing the case in which the labor constraint does not bind, so $w = 1$. Differentiating with respect to ϕ and dividing out the constant

[5] By assuming all ideas are identical from the point of view of consumers, Grossman and Lai [2004] implicitly assume strong return neutrality.

[6] In this case we cannot guarantee that the second-order condition is satisfied, so we must allow the possibility that $\hat{\phi}(\lambda)$ has multiple values, and that for such values of λ the function $\hat{\phi}(\lambda)$ decreases.

$g(\lambda)$, we get the first-order condition for a social optimum:

$$\begin{aligned} FOC(\lambda,\phi) = & \\ & \left[(1/\phi)\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1\right](1/\lambda^2\phi^2)\mu(1/\phi\lambda) \\ & - \int_{1/\phi\lambda}^{\infty} \rho(\overline{\nu}^C - \overline{\nu}^M)\mu(\rho)d\rho \\ & = -\left[(1/\phi)\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1\right](1/\lambda\phi)M'(1/\phi\lambda) \\ & - (\overline{\nu}^C - \overline{\nu}^M)M(1/\phi\lambda). \end{aligned}$$

Divide through by $M(1/\phi\lambda) > 0$. The resulting expression

$$NOC(\lambda,\phi) = \left[(1/\phi)\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1\right]\Upsilon(1/\lambda\phi) - (\overline{\nu}^C - \overline{\nu}^M)$$

has the same qualitative properties as $FOC(\lambda,\phi)$: it has the same zeroes and the same sign on the boundary, and $NOC_\phi(\lambda,\phi) < 0$ is sufficient for a zero to be a local maximum.

We next differentiate with respect to ϕ to find the second-order condition for a social optimum:

$$\begin{aligned} NOC_\phi = & \\ & - \left[(1/\phi)\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1\right](1/\lambda\phi^2)\Upsilon'(1/\lambda\phi) \\ & - \frac{\overline{\nu}^C}{\phi^2}\Upsilon(1/\lambda\phi). \end{aligned}$$

The second term is unambiguously negative. The first term has two factors of interest. We have $(1/\phi)\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1$ representing social surplus of the marginal idea produced; since privately it yields zero profit, it must yield positive social surplus. If the other factor $\Upsilon'(1/\lambda\phi) > 0$, then there is a unique solution to the social optimization problem; if $NOC(\lambda,1) \geq 0$, then that solution is $\hat{\phi}(\lambda) = 1$; otherwise, it is the unique solution to the first-order condition $NOC(\lambda,\phi) = 0$.

In the latter case, we may use the implicit function theorem to compute

$$\begin{aligned} \frac{d\phi}{d\lambda} = -\frac{NOC_\lambda}{NOC_\phi} \propto NOC_\lambda & \\ = -\left[(1/\phi)\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1\right](1/\lambda^2\phi)\Upsilon'(1/\lambda\phi), & \end{aligned}$$

which has the opposite sign to $\Upsilon'(1/\lambda\phi)$. This covers the second half of the proposition when the labor constraint does not bind.

If the labor constraint does bind, increasing ϕ only increases the wage rate. Hence, if the social optimum is to allow the labor constraint to bind, ϕ must be chosen as small as possible subject to the constraint of full labor utilization and $w = 1$. Consequently, concavity of welfare in the interior implies a unique optimal choice of ϕ. This establishes a unique optimal policy function $\hat{\phi}(\lambda)$ when $\Upsilon'(1/\lambda\widehat{\phi}(\lambda)) \geq 0$.

Finally, we turn to the first half of the proposition. For fixed ϕ and all large enough λ, we can assume that either $\Upsilon'(1/\lambda\phi) > 0$ or $\Upsilon'(1/\lambda\phi) < 0$.

In either case, $\Upsilon(\rho)$ must have a (possibly infinite) limit as $\rho \to 0$. Observe that $\Upsilon(\rho) \equiv -\rho M'(\rho)/M(\rho)$ and that $M(\rho)$ is nonincreasing. Suppose first that $-\rho M'(\rho)$ does not converge to infinity. If it is bounded away from zero, $M(\rho) \to \infty$, implying $\Upsilon(0) = 0$. If it is not bounded away from zero, since $M(\rho)$ is bounded away from zero, again, $\Upsilon(0) = 0$. Hence, either $-\rho M'(\rho) \to \infty$ or $\Upsilon(0) = 0$. The latter case implies near $0, \Upsilon'(\rho) > 0$, so fix $\phi = 1$ and examine

$$NOC(\lambda, 1) = \left[\overline{\nu}^M - 1\right] \Upsilon(1/\lambda) - (\overline{\nu}^C - \overline{\nu}^M).$$

Since $\Upsilon(0) = 0$ for λ sufficiently large $NOC(\lambda, 1) < 0$, implying $\hat{\phi}(\lambda) < 1$. It now follows from the first part of the proof that $\hat{\phi}(\lambda)$ is strictly decreasing.

Finally, then, suppose $-\rho M'(\rho) \to \infty$. The demand for labor is

$$L^D = g(\lambda) \int_{1/\phi\lambda}^{\infty} \mu(\rho) d\rho.$$

Differentiating with respect to λ yields

$$D_\lambda L^D = g'(\lambda) \int_{1/\phi\lambda}^{\infty} \mu(\rho) d\rho + g(\lambda)(1/\phi\lambda^2)\mu(1/\phi\lambda).$$

Labor supply is λL, so if $D_\lambda L^D \geq L + \epsilon$ for all sufficiently large λ, the labor constraint must eventually bind. But $-\rho M'(\rho) = \rho^2 \mu(\rho) \to \infty$ as $\rho \to 0$. So, for ϕ bounded away from zero, $D_\lambda L^D \to \infty$. Hence, in this case the labor constraint must bind. ∎

Three Implications. Theorists of innovation and growth due to aggregate increasing returns often insist that the latter are due to massive externalities in the creation of new ideas. In our framework this requires $g(\lambda)/\lambda$ to be increasing in λ; that is, the number of available ideas increases more rapidly than the scale of the economy. This leads to a number of extravagant predictions, which we will compare to the empirical evidence later. Here it suffices to remark that, as a matter of theory, $g(\lambda)/\lambda$ increasing in λ implies that the labor constraint must bind eventually; when it does the private return from the marginal idea adopted increases rather than decreases. While this does not imply that ideas with a higher social value will be produced, it certainly does imply that the optimal length of protection ϕ is strictly decreasing in λ.

That intellectual property protection can drive up the wage rate for the relevant supply of highly skilled labor is empirically relevant for policy. Lobbyist groups, especially in the entertainment and pharmaceutical industries, often point to the high cost of producing new goods as a reason for strong IP protection. Examination of the balance sheets of either movie production or pharmaceutical companies shows that much of this high cost is due to the earnings of the "highly skilled" labor these, and other IP protected industries, employ. Consider the entertainment industry: costs are high here because a few"stars" earn large amounts of money. Since the opportunity

cost of these people is generally quite small, an important effect of reducing copyright protection will simply be to lower the rents earned by these "stars," and consequently to reduce the cost of producing movies of a given quality. In such industries, the marginal workers are paid close to their opportunity cost and so stand to lose little through reduced copyright protection. A similar, even if admittedly less straightforward, argument can be applied to the drug industry with respect to the wages of medical researchers in relation to that of production workers.

Other, comparative static results also follow from the previous analysis. Any policy or technological change increasing the marginal cost of the skilled labor needed to introduce new ideas is equivalent to decreasing λ, so generally such changes increase the socially optimal length of protection. This observation is relevant when confronting policies that increase the protection of selected groups of skilled workers from foreign labor competition (for example, by restricting immigration or penalizing outsourcing); in this case, protection in the labor market induces additional IP protection and rent-seeking in the product market. Conversely, technological improvements (such as the increasing power and reduced cost of computers or the reconstruction of the DNA code) which reduce the size of the initial indivisibility h are tantamount to increasing λ; hence, they should engender a reduction of the socially optimal length of protection.

The Production Function Approach. By way of contrast, Grossman and Lai [2002, 2004] adopt a production function approach where ideas are homogeneous and the total number of ideas is $Q = F(H, L)$, where H is a fixed amount of human capital, L is the labor input, and F is a constant returns to scale concave production function. Since human capital serves in this model only to absorb the rents from ideas, we may as well write $Q = f(L)$, where f is a diminishing returns production function. While we do not think that ideas are like automobiles—more or less perfect substitutes stamped off a production line—we can construct a total monopoly revenue schedule corresponding to Grossman and Lai's production function and by doing so give an interpretation of their assumption in our framework. Observe that when $w = 1$ the total labor cost of producing Q ideas is $f^{-1}(Q)$, and the corresponding marginal cost is $1/f'(L)$. Since all ideas are equally valuable, we may as well suppose they generate revenue 1, so in our terminology, the private return of the idea produced by the Lth unit of labor input is revenue divided by the cost of producing the idea; that is, $\rho = f'(L)$. The total revenue to ideas with private return ρ or better is then the total number of ideas produced by the corresponding amount of labor

$$M(\rho) = f([f']^{-1}(\rho)).$$

>From this we can easily derive that the elasticity Υ of M is the same as Grossman and Lai's elasticity of research output with respect to labor.[7] In

[7] In their notation, this elasticity is γ.

the Cobb-Douglas case, which is the benchmark case studied by Grossman and Lai, $f(L) = L^{\alpha}$, and so $M(\rho) = \alpha^{\alpha/(1-\alpha)}\rho^{-\alpha/(1-\alpha)}$, which is to say that M has constant elasticity. Also of interest is their CES case, where $f(L) = (a + L^{\beta})^{1/\beta}$for $\beta \leq 1$. We have $f'(L) = (aL^{-\beta} + 1)^{(1-\beta)/\beta}$,

$$M(\rho) = a^{1/\beta}\left(\frac{1}{1-\rho^{-\beta/(1-\beta)}}\right)^{1/\beta}$$

$$\Upsilon(\rho) = \frac{1}{1-\beta}\frac{1}{\left(\rho^{\beta/(1-\beta)}-1\right)}.$$

For $\beta < 0$, $M(\rho)$ is defined for $\rho \leq 1$, with $M(1) = 0$, and elasticity is increasing in ρ. The case $0 < \beta \leq 1$ implies that even with no labor input, $a^{1/\beta}$ ideas will be produced. This is not ridiculous: even in the absence of any effort, some ideas with a positive return may be discovered by accident. Also in this case $M(\rho)$ is defined for $\rho > 1$ with $M(1) = \infty$, that is, the revenue generated by ideas goes to infinity, even before all ideas of positive quality are exhausted. Here we have decreasing elasticity.

4. The Elasticity of Total Monopoly Revenue

The question then is, Which assumption is most appropriate on the elasticity of the total monopoly revenue function? We address it here from both a theoretical and an empirical perspective. First we look at distributions that are interesting either because they satisfy intuitive properties or because they allow for explicit computations and are widely used in practical examples. With the exception of the Pareto, we find that the elasticity of $M(\rho)$ is increasing. Next we look at empirical distributions of income/revenue for different kinds of inventive activity and find that they exhibit rapidly rising elasticity. Finally, we derive the implications of the constant elasticity assumption for the equilibrium demand of labor in the idea sector, as the scale of the market increases. We show that assumptions of constant or decreasing elasticity is grossly at odds with any available data, and probably also with common sense.

4.1. **Theoretical Analysis.** First, we consider from a theoretical point of view what

$$M(\rho) = \int_{\rho}^{\infty} \rho'\mu(\rho')d\rho'$$

might look like. Under the plausible assumption that there are ideas so bad that they have a negative private return, we expect $\mu(0)$ to be strictly positive, and finite. This implies that $M(0)$ is finite, and $M'(0) = 0$, so $\lim_{\rho\to 0}\Upsilon(\rho) = 0$. Since $\Upsilon(\rho) \geq 0$, this means $\Upsilon'(0) \geq 0$, that is, the increasing elasticity case. In other words, theoretical considerations alone suggest that the function $M(\rho)$ is finite and flat at $\rho = 0$ and has increasing elasticity there. We should not expect situations such as that implied by the CES production function for $0 < \beta \leq 1$ in which elasticity is globally decreasing, or even constant.

4.1.1. *The Pareto Distribution.* If $\mu(\rho)$ is Pareto, then $M(\rho) = \rho^{-\zeta}$, which corresponds, up to a scale factor, to the functional form implied by the Cobb-Douglas production function. Since the Pareto density goes to infinity for finite ρ, as we have observed, this is not a terribly good global model of the distribution of ideas, and we would not expect it to hold for ρ close to zero. Nevertheless, it can certainly be argued that we are still in the upper tail of the distribution of quality of ideas. And there is certainly a great deal of economic data that appears to have Pareto upper tails, so it is possible in principle that we are still experiencing values of ρ for which $\Upsilon'(\rho) \leq 0$. Or, perhaps, the tails are even thicker than Pareto, with a lump of ideas with zero cost, as is the case for the CES production function with $0 < \beta \leq 1$. This is an empirical issue, and we will address it shortly.

Next we quickly run through various functional forms for $\mu(\rho)$ that might be useful to model the distribution of ideas and see what they imply for $M(\rho)$ and $\Upsilon(\rho)$.

4.1.2. *The Exponential, Normal, and Lognormal Distributions.* Consider first the Exponential, where $\mu(\rho) = \zeta^3 e^{-\zeta\rho}$. Then $M(\rho) = \zeta(\zeta\rho + 1)e^{-\zeta\rho}$ and

$$\Upsilon(\rho) = (\zeta\rho)^2/(1 + \zeta\rho),$$

which is increasing in ρ. For the Normal distribution, say $\mu(\rho) = e^{-\zeta\rho^2}$, we have that $M(\rho) = (2\zeta)^{-1}e^{-\zeta\rho^2}$ and $\Upsilon(\rho) = 2\zeta\rho^2$, which is also increasing for all values of $\rho > 0$. Similar calculations yield equivalent results for a Lognormal distribution of private returns. Should we restrict ourselves, as it is reasonable, to positive and large values of ρ, the results do not change, in fact, they are strengthened. For the Normal distribution, $\Upsilon(\rho)$ is increasing in ρ over the whole interval $(0, \infty)$, constant at $\rho = 0$, and decreasing for negative values of ρ. For the Exponential, $\Upsilon(\rho)$ is increasing for all $\rho > 0$ and for $\rho < -2/\zeta$ as well.

4.1.3. *The Truncated Pareto Distribution.* In principle, a Pareto distribution with a finite upper bound is appealing, and it allows for explicit computations. It amounts to assuming that private returns are distributed Pareto and there is an idea of highest possible quality, so $\mu(\rho) = \rho^{-\zeta}$ for $\rho \leq \overline{\rho}$, and $\mu(\rho) = 0$ for $\rho > \overline{\rho}$, with $\zeta < 2$. (But notice that for values of $\zeta < 0$ this says that the frequency of ideas increases with their quality, and then suddenly drops to zero after the maximum value $\overline{\rho}$ is reached, which does not make much practical sense, so assume $\zeta > 0$.) In this case, again, elasticity rather than being constant is increasing. In fact, we have

$$M(\rho) = \frac{1}{2-\zeta}\left[\overline{\rho}^{2-\zeta} - \rho^{2-\zeta}\right]$$

$$\Upsilon(\rho) = \frac{2-\zeta}{(\overline{\rho}/\rho)^{2-\zeta} - 1}$$

$$\Upsilon'(\rho) = \frac{(2-\zeta)^2\,(\overline{\rho})^{2-\zeta}\rho^{\zeta-3}}{\left((\overline{\rho}/\rho)^{2-\zeta} - 1\right)^2} > 0.$$

In the case $\zeta = 1$, we see that $M(\rho)$ is linear, and we can explicitly solve the NOC:

$$NOC(\lambda, \phi) = \left[(1/\phi)\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1\right]\frac{1}{\lambda\phi\overline{\rho} - 1} - (\overline{\nu}^C - \overline{\nu}^M) = 0.$$

We rewrite the NOC as

$$\lambda\overline{\rho}(\overline{\nu}^C - \overline{\nu}^M)\phi^2 + \phi - \overline{\nu}^C = 0$$

and then find the positive root

$$\phi = \frac{\sqrt{1 + 4\lambda\overline{\rho}(\overline{\nu}^C - \overline{\nu}^M)\overline{\nu}^C} - 1}{2\lambda\overline{\rho}(\overline{\nu}^C - \overline{\nu}^M)}.$$

If we normalize $\overline{\rho} = 1$, then for large λ this can be approximated by

$$\phi \approx \left(\frac{\overline{\nu}^M}{\overline{\nu}^C - \overline{\nu}^M}\right)^{1/2}\lambda^{-1/2},$$

which implies an elasticity of protection with respect to scale of market of 1/2. That is, quadrupling the scale of the market implies that protection should be reduced by about a factor of two.

Returning to the case of general ζ, consider the NOC for $\phi = 1$:

$$NOC(\lambda, 1) = \left[\overline{\nu}^M - 1\right]\Upsilon(1/\lambda) - (\overline{\nu}^C - \overline{\nu}^M).$$

If $\lambda > 1/\overline{\rho}$ holds, that is, the size of the economy is large enough, then $NOC(\lambda, 1) < 0$, implying an interior solution. This shows that the size of the economy and the upper bound on the achievable private return on ideas play a similar role in our model; a small economy where very highly profitable ideas are available is equivalent, from the viewpoint of optimal protection, to a large one in which only not-so-profitable ideas are available.

Now, let $\overline{\rho}\lambda \leq 1$ hold, so that $NOC(\lambda, 1) > 0$, so $\phi = 1$ is optimal. This is an economy either of very small size or in which the private return from new ideas is very low. In this case even complete monopoly cannot help: with $\phi = 1$ the marginal idea produced is $\rho = 1/\lambda$, implying that $\rho \geq \overline{\rho}$ holds, that is, no idea is ever implemented.

Finally, notice that, in general, labor demand is

$$L^D = g(\lambda)\int_{1/\lambda\phi}^{\overline{\rho}} \rho^{-\zeta} d\rho = \frac{g(\lambda)}{\zeta - 1}[(\phi\lambda)^{\zeta-1} - \overline{\rho}^{\zeta-1}]$$

while labor supply is $L^S = L\lambda$. Suppose the labor constraint is binding; equilibrium is achieved either by lowering labor demand through the wage rate or through the protection level ϕ. We know social optimality requires keeping the wage rate at 1 and equilibrating the labor market using the protection level ϕ. This gives the optimal protection level of

$$\phi = \left[\left(\frac{1}{\lambda\overline{\rho}}\right)^{\zeta-1} + \frac{(\zeta-1)L}{g(\lambda)\lambda^{\zeta-2}}\right]^{1/\zeta-1}$$

which, once again, is strictly decreasing in λ and $\overline{\rho}$.

4.2. **Empirical Analysis of Total Monopoly Revenue.** Up until now we have been thinking of ideas as empty boxes to be filled in by individuals. From an empirical perspective, it is more useful to think of each individual being associated with her own ideas, that is, to think of ω as indexing individuals or, rather, individuals' private return from investing in ideas. Then $h(\omega)$ becomes equal to the cost of person ω's time, or if labor is equally productive in the nonidea sector, $h(\omega)$ would be the same for all individuals. This is a strong, but not incredible, assumption that is more or less the opposite of the production function approach; we will adopt it to start with and consider its weaknesses later. That is, set $h(\omega) = 1$; then, any individual with an idea of expected value higher than 1 would try to implement it. We then identify individuals with their private returns ρ and think of them as equivalent to the expected value of their ideas, with the latter being drawn from an underlying distribution $\mu(\rho)$ satisfying the restrictions discussed earlier. We are interested in the shape of $\mu(\rho)$ as this would allow us to compute the elasticity of $M(\rho)$ at the "cutoff idea-individual" $\underline{\rho} = 1/\phi\lambda$.

An issue arises when going to the data. In our model, $v(\omega, z) = p(\omega)u(\omega, z)$, where $1-p(\omega)$ is the probability that *ex post*, after the project is paid for, the idea turns out to have no value. Unfortunately, we rarely, if ever, observe the price $v_z(\omega, z) = p(\omega)u_z(\omega, z)$, but rather just $u_z(\omega, z)$. As long as $p(\omega)$ does not depend on ω, this means that we will overestimate returns and revenues, but will correctly find elasticities. So we must assume not only that $h(\omega)$ is constant, but $p(\omega)$ is as well. We have also assumed that there is only one "successful" *ex post* outcome; if there is a large variation in *ex post* outcomes, this too will pose a problem for data analysis. Movie revenue data, for example, would pose a huge problem in this regard, because of the high degree of *ex ante* uncertainty about how much a given film is likely to earn.

4.2.1. *Personal Income Distribution.* Our first attempt uses data for the U.S. income distribution. That is, we make the further assumption that the distribution of income among creative individuals is the same as for the population at large. This is probably incorrect when it comes to levels: creative individuals are likely to be concentrated in the upper tail of the distribution of personal income. However, it is a plausible assumption about the shape of the distribution of income from creative activity. In any case, to the extent that the largest share of personal income is due to labor effort and people use their creativity in accumulating skills and choosing an occupation, this is a reasonable starting point. Current Population Survey data on income from 2001 are shown in Table 1. Abusing notation slightly, let ρ denote personal income here; the corresponding $M(\rho)$ is plotted in Figure 1.

Table 1

Income in USD 1,000s	Population
0-10	9.0
10-20	6.9
20-30	13.3
30-40	12.4
40-50	15.4
50-75	18.4
75-100	10.8
100+	13.8

Sophisticated econometric and statistical knowledge is not required to see that this curve is well fit by a straight line and poorly by a Pareto distribution. The U.S. cumulative distribution of personal income, clearly, has increasing elasticity.

4.2.2. *Revenue from Authorship of Fiction Books.* We now examine a particular category of creative individuals: authors of fiction books. Ideally, one would like to observe revenues for various books for each author, to account for the possible *ex ante* uncertainty about *ex post* sales. Such data are not available; hence, we proceed with what is. Although we do not have data on lifetime income of individual authors, we do have data on the revenue generated by individual book sales. Looking directly at revenues from book sales is interesting because, as noticed, authors may earn above average income, and indeed may be largely in the upper tail of the income distribution—which is not measured at all by the income data we reported earlier. It might well be that while the bulk of the distribution of personal income is linear, it is Pareto in the upper tail, and the income of authors might turn out to be distributed like the upper tail of overall personal income. Assuming $h(\omega)$ constant in this case means that all those that "give it a try" at writing fiction have equivalent opportunity costs. We ignore the problem of initially deciding to be an author and focus on the decision to continue writing books after the first. For new authors, there is an option value to producing the first book—since if it is a failure, there is no reason to continue as an author. Our static model does not capture this type of option value. But if we assume that the cost of writing the first book is small relative to the lifetime cost of being an author, this will not matter much. We also ignore the fact that it is costly to produce books once they are written, which is largely irrelevant to our ends to the extent that the cost of producing each copy of a book is independent of the number of copies produced and sold. Suppose the following:

- All authors take the same amount of time to produce a novel and have the same opportunity cost, so $h(\omega)$ is constant.
- Authors earn all their income from the sale of their novels.

- Expected revenues from the sale of a "successful" book are perfectly anticipated, and the probability of failure does not depend upon ω.

Then income per unit of time taken to produce a book is $r = \lambda\phi\rho$ and, given current copyright laws, one can safely set $\phi = 1$ in what follows. We can compute the aggregate income of all authors who earn at least a given amount, $M^r(r)$, and of course $M(\rho) = (1/\lambda)M^r(\rho/\lambda)$ has the same elasticity. We gathered data on revenues for 1223 and 1235 fiction books published in September 2003 and September 2004, respectively. Figure 2 shows $M(\rho)$ for September 2003. Figure 3 shows a plot on logarithmic axes, including a closeup to illustrate more clearly the increasing nature of the elasticity on both ordinary and logarithmic axes. Figure 4 shows the September 2004 data.

Examining these plots, four things stand out.

First, the graphs are very similar both in shape and absolute values. This suggests that (i) the pattern reported here is quite robust, and (ii) even for successful books most sales take place in the first few months after publication.[8]

Second, for less successful books the $M(\rho)$ function is nearly linear, and overall the function exhibits increasing elasticity—a fact that can be seen more clearly in the logarithmic plots.

The third striking feature is the discontinuity between roughly \$150,000 and \$300,000 in revenue.[9] This is most pronounced in 2004, where 5 books out of 1235 generate nearly 25% of the revenue.[10] These books appear to be predominately by "big name" authors,[11] who are largely irrelevant for optimal copyright policy: the relevant part of the $M(\rho)$ function is the part near the cutoff—that is, for marginal, not inframarginal, books.

The fourth is how low the indivisibility $wh(\rho)$ may be for writing and publishing fiction; in 2003, 1181 books, out of a total of 1223, earned \$50,000 or less (corresponding to total revenue of approximately \$300,000). These books accounted for 50% of total revenue, that is, \$6M out of \$12M. The numbers for 2004 are similar. In the September 2003 data, 984 books earned less than \$10,000; hence, our estimate of the marginal author's opportunity cost $wh(\underline{\rho})$ should be placed at \$60,000 or less. In light of our earlier discussion, such a conclusion is valid only if those authors and their publishers, correctly predicted actual sales. Certainly books earning \$300,000

[8]Data for the months of March 2003 and March 2004 confirm this is indeed the case.

[9]The sales data are from a single distributor, Ingram, constituting about 1/6 of the book market, so total revenues would be about six times this number.

[10]This is broadly consistent with other data on books revenues: Leibowitz and Margolis [2003] report that less than 200 out of 25,000 titles account for roughly 2/3 of all book revenues. This is considerably more concentrated than we find in our data—but certainly reflects a strong discontinuity.

[11]Of the five books earning above \$300,000 in 2003, three are by authors with previous best-sellers, one by a well-established author who had not previously been on the New York Times best-seller list, and one was by a first-time author.

in revenue are paying the opportunity costs of the authors, and it is hard to imagine that 1235 books were published during September 2004 hoping to be one of the 5 that generate more than $300,000 in revenue.

What are the implications of these data for the optimal ϕ, and how does this compare to actual copyright protection? Since the Sonny Bono copyright extension act of 1998, copyright protection in the United States is life of the author plus 70 years, or 90 years for works without an author. If we take the remaining life of an author to be roughly 35 years, this would mean 105 years of protection.[12] If the flow of sales is constant over time, for interest rate r and copyright length T, the corresponding value of $\phi = 1 - e^{-rT}$. The value of ϕ for various real interest rates is given in Table 2.

Table 2

r	ϕ
0.01	0.650
0.02	0.878
0.03	0.957
0.04	0.995
0.05	0.999

Since no reasonable estimate of the real interest rate is below 0.2, this means that the current ϕ is fairly close to 1. Moreover, the flow of sales is far from uniform over time; in our data for books published in September 2003, during the four months of 2003 revenues were 2.4 times the revenue during the 10 months of 2004, meaning that per month sales fell by a factor of 6. This is consistent with the general claim that the most significant book sales occur within three months of publication.[13]

The elasticity $\Upsilon(\rho)$ of $M(\rho)$ for the 2003 data is shown in Figure 5. From the NOC we can compute the optimal protection as a function of that elasticity

$$\hat{\phi} = \left(\frac{1}{\overline{\nu}^C} + \frac{\overline{\nu}^C - \overline{\nu}^M}{\overline{\nu}^C} \frac{(1+\Upsilon)}{\Upsilon} \right)^{-1}.$$

In the case of linear demand, $\overline{\nu}^C = 2, \overline{\nu}^M = 3/2.$. Table 3 reports the corresponding lengths of optimal copyright for various real interest rates and elasticities in the range in Figure 5.

[12] Akerloff et al. [2002] use an estimate of 30 additional years of life and a 7% real interest rate.

[13] Our findings are somewhat different from those of Leibowitz and Margolis [2003], who argue that for popular books, revenues are somewhat more spread out over time. In our data, the top 9 books (those above the discontinuity) earned 3.3 times more revenue in 2003 as in 2004. This has the implication that effective protection is slightly greater for inframarginal books than marginal books—exactly the opposite of what is socially desirable.

Table 3

Υ	$\hat{\phi}$	$r = 0.2$	$r = 0.4$
0.05	0.13	7	4
0.10	0.24	14	7
0.15	0.33	20	10
0.20	0.40	26	13
0.30	0.51	36	18
0.40	0.60	46	23

Two facts stand out. First, optimal length of protection is less than 1—meaning that given the elasticity is increasing, optimal copyright protection should decline with the size of the market. Second, optimal copyright length is much less than actual copyright length; since the actual cutoff value of ρ in the data is quite small, even an elasticity of 0.05 may be a tremendous overestimate of the actual elasticity on the margin. Certainly it is hard to justify even 7 years of copyright based on these data. This evidence argues quite strongly that copyright protection is determined by rent-seeking and not as an optimal policy: the only real beneficiaries from copyright extensions that increase the effective protection from, say, 0.98 to 0.99 are not potential authors, but holders of current copyrights that are about to expire. Transferring money to existing copyright holders, of course, serves no useful economic purpose.

4.2.3. *Patent Values.* A similar analysis of the value of patents is possible—with the reservation that it is less likely for patents that *ex post* value can be anticipated *ex ante*. However, it may be that the major uncertainty in patenting is not how much will be earned if the project is successful, but rather, how likely the patent is to succeed, in which case our model applies. If we disaggregate by industry, it is at least plausible that the fixed cost of the innovation is not systematically related to the *ex post* private return. We use *ex post* data on the value of patents from Lanjouw [1993] for four German industries—estimated from patent renewal rates and data on the cost of renewal. We graph the corresponding $M(\rho)$curves in Figure 6.

We compute the elasticities of the linear spline at the midpoint of the intervals in Table 4.

Table 4

Computers	Pharmaceuticals	Textiles	Engines
..22 [.17]	.14 [.12]	.19 [.15]	.32 [.23]
..74 [.40]	.53 [.33]	.66 [.38]	.95 [.45]
..93 [.30]	.75 [.30]	.88 [.31]	1.12 [.32]
3.76 [.60]	2.35 [.48]	2.42 [.44]	3.04 [.42]
2.73 [.12]	2.81 [.16]	3.02 [.14]	3.37 [.12]

As can be seen, in no case are the tails similar to that of a Pareto distribution—the curves fall far too close to zero. Moreover, with the (irrelevant) exception of the highest category of ρ for computers, they exhibit increasing elasticity over every portion of their range. Table 4 also reports in square brackets $-\rho M'(\rho)$. This is relevant because the same ϕ (i.e., 20 years) applies across sectors. Hence, the relevant distribution is $M(\rho) = \sum_i M_i(\rho)$, where i indexes the various industries subject to patenting. Unfortunately, the fact that each $M_i(\rho)$ function has increasing elasticity does not imply that this is true in the aggregate. Hence, it is of interest to examine $-\rho M_i'(\rho)$. If this is increasing, then the corresponding elasticity is increasing as well, and increasing $-\rho M_i'(\rho)$ is a condition that does aggregate. While not increasing in (virtually) every case as is the elasticity, $-\rho M_i'(\rho)$ is increasing at the lower end (near the existing threshold) and increases in most cases. This means that these results can be expected to aggregate over the different industries.

One thing that should be clear about this analysis is that existing data are not ideally suited to examining the $M(\rho)$ function. In particular, estimate of the anticipated and unanticipated components of return, and of the fixed cost $h(\omega)$, would greatly improve the analysis.

Our findings for patents appear to accord well with existing, and more elaborate, literature on the same subject. To name but a few recent studies, Harhoff, Scherer, and Vopel [1997] use a data set of full-term patents applied for in 1977 and held by West German and U.S. residents. They compare the performances of various empirical distributions, including Pareto's, to fit the data and find that a two-parameter log normal distribution provides the best fit. Silverberg and Verspagen [2004] use a variety of data sources (European and U.S. Patent Offices, as well as a data set on CT scanners) and different measures of ρ (citations and monetary values) in their estimation. They find that, while the overall distributions are well approximated by exponential ones, it is the *upper tail* that is better captured by a Pareto distribution. As our concern here is with the shape of the $\mu(\rho)$ near the lower cutoff value $\underline{\rho}$, this is supportive of our claim. The older literature on the value of patents, stemming from the path-breaking paper of Pakes [1986] (see Hall, Jaffe, and Tratjenberg [2004] for a recent update and new results), seems to find, almost always that the appropriate distribution is a log normal or an exponential, for both of which the elasticity of the total revenue function is increasing. Interestingly, Sampat and Ziedonis [2002] find that citations are not a good predictor of revenues earned from licensed patents.

As in the case of copyright, the elasticities we observe argue that, if demand is linear, optimal protection $\hat{\phi}$ should be considerably less than one, implying in addition that optimal protection should decline with the scale of the market. Ignoring the fact that patents are likely to generate considerably more revenue early in their life than late in their life, and referring back to Table 3, the middle ground elasticity of .15 and a real interest rate of 4% imply an optimal patent length of 10 years, while at the high end with an

elasticity of .4 and a real interest rate of 2%, the optimal term would be 46 years. The existing U.S. term of 20 years lies in between. We believe that the higher interest rate and lower elasticity are more reasonable. When we combine this with the fact that an idea generates most revenue early in its life, it suggests to us that current patent terms are too long. However, the situation is clearly not as extreme as in the case of copyright.

4.2.4. *Analysis of Labor Demand: Theory.* Another route to determining whether elasticity is increasing or decreasing is to study its implications for the labor demand in the idea sector. This is

$$L^D(\lambda) = g(\lambda) \int_{1/\phi\lambda}^{\infty} \mu(\rho) d\rho = g(\lambda)\ell(1/\phi\lambda).$$

Letting El denote the elasticity operator, its elasticity is

$$El_\lambda[L^D(\lambda)] = El_\lambda[g(\lambda)] - El_\rho[\ell(\rho)].$$

Depending on which assumptions one makes about $g(\lambda)$, the first factor ranges from zero to any large positive number. For example, Grossman and Lai [2004] identify $g(\lambda)$ with aggregate human capital H in their model and assume this is constant relative to market size; hence, $El_\lambda[g(\lambda)] = 0$. As pointed out in Section 3, in models of growth and innovation due to externalities, such as Grossman and Helpman [1991, 1994, 1995], or Romer [1990], $g(\lambda)$ increases faster than λ; hence, $El_\lambda[g(\lambda)] > 1$. A benchmark case is that in which each individual draws her own ideas from the same urn, with or without replacement. If sampling is without replacement, and each person draws the same number of ideas for each characteristic ω, then $g(\lambda) = \lambda$ and $El_\lambda[g(\lambda)] = 1$; if sampling is with replacement then, in general, we would have $El_\lambda[g(\lambda)] \leq 1$.

As for the second factor, notice first that the demand for labor is

$$\ell(\rho) = \int_{\rho}^{\infty} -[DM(\rho')/\rho']d\rho'.$$

Now, assume that $M(\rho) = \rho^{-\zeta}$, which is the constant elasticity case. Then

$$\ell(\rho) = \frac{(\zeta+1)(\rho)^{-1-\zeta}}{\zeta+2},$$

and if $El(g(\lambda)) \geq 1$,

$$El(L^D(\lambda)) = El(g(\lambda)) + \zeta + 1.$$

Notice that, when $El(g(\lambda)) > 1$, the elasticity of labor demand is predicted to be *substantially* larger than two, or the elasticity of per capita labor demand is greater than one. This implies that in the data we should observe that, as the size of the economy grows, the *share* of workers in the idea sector grows more than proportionally. This prediction is reinforced when the elasticity of the total monopoly revenue function is decreasing, as we show next.

Proposition. *Consider two aggregate monopoly revenue functions M_1, M_2 that have the same value $M_1(\rho) = M_2(\rho)$ and derivative $DM_1(\rho) = DM_2(\rho)$ (hence, elasticity $\Upsilon_1(\rho) = \Upsilon_2(\rho)$) at ρ. If $D\Upsilon_1(\rho') < D\Upsilon_2(\rho')$ for $\rho' \geq \rho$. Then*

(1) Labor demand associated to M_1 is smaller than the one associated to M_2; that is,
$$\int_\rho^\infty -[DM_1(\rho')/\rho']d\rho' < \int_\rho^\infty -[DM_2(\rho')/\rho']d\rho'.$$
(2) The elasticity of labor demand associated to M_1 is greater than the elasticity of labor demand from M_2; that is, $El[\ell_1(\rho)] > El[\ell_2(\rho)]$.
(3) As the elasticity of total revenue goes from increasing, to constant, to decreasing, the elasticity of the associated labor demand functions increase monotonically.

Proof. **Step 1:** $M_1(\rho') > M_2(\rho')$.

Here and in what follows, $\rho' \geq \rho$ holds. Then, $D\Upsilon_1(\rho) - D\Upsilon_2(\rho) < 0$ by assumption. Moreover,
$$D\Upsilon(\rho) = D[-\rho DM(\rho)/M(\rho)]$$
$$= \frac{1}{\rho}[\Upsilon(\rho) + \Upsilon^2(\rho)] - \rho D^2 M(\rho)/M(\rho),$$
so $D^2M_2(\rho) - D^2M_1(\rho) = (M(\rho)/\rho)[D\Upsilon_1(\rho) - D\Upsilon_2(\rho)] < 0$, where $M(\rho)$ is the common value of M_1 and M_2 at ρ. Then, for ρ' near ρ we have
$$M_1(\rho') - M_2(\rho') \approx (1/2)[D^2M_1(\rho) - D^2M_2(\rho)](\rho' - \rho)^2 > 0.$$
Moreover, if $M_1(\rho'') - M_2(\rho'') < 0$ for some larger ρ'', then $M_1(\rho') - M_2(\rho') = 0$ for some $\rho'' > \rho' > \rho$, since both functions are continuous. Let $\hat{\rho}'$ be the smallest such ρ', that is, the first point to the right of ρ where M_1 and M_2 cross. Then $\Upsilon(\hat{\rho}') = -\rho' DM(\hat{\rho}')/M(\hat{\rho}')$ and the assumption that $\Upsilon_1(\hat{\rho}') < \Upsilon_2(\hat{\rho}')$ imply $DM_1(\hat{\rho}') > DM_2(\hat{\rho}')$; that is, M_1 crosses M_2 from below, which is impossible since to the left of $\hat{\rho}'$ we already know that $M_1 > M_2$.

Step 2: $\int_\rho^\infty -[DM_1(\rho')/\rho']d\rho' < \int_\rho^\infty -[DM_2(\rho')/\rho']d\rho'$.

Recall that $M(\infty) = 0$. Integration by parts gives
$$\int_\rho^\infty -[DM(\rho')/\rho']d\rho' = -M(\rho')/\rho'|_\rho^\infty - \int_\rho^\infty M(\rho')/(\rho')^2 d\rho'$$
$$= M(\rho)/\rho - \int_\rho^\infty M(\rho')/(\rho')^2 d\rho'$$
from which
$$\int_\rho^\infty -[DM_1(\rho')/\rho']d\rho' - \int_\rho^\infty -[DM_2(\rho')/\rho']d\rho'$$
$$= -\int_\rho^\infty [M_1(\rho') - M_2(\rho')]/(\rho')^2 d\rho' < 0.$$

Step 3: $El[\ell_1(\rho)] > El[\ell_2(\rho)]$.
Because

$$El[\ell(\rho)] = El[\int_{\rho}^{\infty} -[DM(\rho')/\rho']d\rho']$$
$$= \frac{-\rho DM(\rho)/\rho}{\int_{\rho}^{\infty} -[DM(\rho')/\rho']d\rho'}$$
$$= \frac{-DM(\rho)}{\int_{\rho}^{\infty} -[DM(\rho')/\rho']d\rho'}.$$

$El[\ell_1(\rho)]$ and $El[\ell_2(\rho)]$ have the same numerator, and, because of Step 2, the first has a smaller denominator. Hence, the conclusion. ∎

In plain words: if a constant elasticity revenue function implies an elasticity of labor supply with respect to market size larger than one, then a revenue function with decreasing elasticity would imply an even larger elasticity of labor supply. Playing this backward: should the empirical elasticity of per capita labor supply with respect to changes in market size be smaller than one, then the associated total revenue function would have to have increasing elasticity, which is our claim. It is important to stress that, on the basis of the algebra above, even if per capita labor in the idea sector grows faster than the scale of market, this is consistent with increasing elasticity of total monopoly revenue. If per capita labor grows more slowly, we can rule out decreasing elasticity, but not the other way around. This is because $El_{\lambda}[g(\lambda)]$ can be large, which is independent of the elasticity of monopoly revenue.

Common sense is probably enough to reject the hypothesis that the elasticity of the total revenue function is constant or decreasing, as the latter implies that when the market size doubles the share of population dedicated to creative activity more than doubles. During the last century, the effective size of the world market has increased, roughly, by a factor of 100. This implies that, at least in the United States, by now everybody should be working in the idea sector, which, alas, we have not yet achieved.

4.2.5. *Analysis of Labor Demand: Copyright.* With this background, let us look at some data. First we consider the copyright time series.[14] Here we must assume that the distribution $M(\rho)$ is time invariant—for example, it is not the case that all good ideas have been used up. We also must assume that ϕ is either constant or increasing over time—as, in fact, it is. We measure the scale of the market by the size of the literate population,[15] and the amount of labor in the sector by the number of copyright registrations. The relevant annual growth rates for the United States are reported, by decade, in Figure 7. If elasticity of total monopoly revenue is, in fact, constant or decreasing, we expect to see per capita copyright growing more rapidly

[14]The data sources are described in detail in the Appendix.

[15]The literacy adjustment makes little difference; in 1870, when the copyright registration data begin, the literacy rate is already 80%, climbing to 92.3% by 1910.

than population. This is, in fact, the case prior to 1900, and after 1970. This means that, for those periods, we cannot rule out the possibility that elasticity was constant or decreasing. For the pre-1900 period one must notice that copyright registration only begins in 1870, so the huge initial increase in registrations is unlikely to reflect a corresponding increase in the actual output of literary works. In particular, it is important to realize that in 1891 it became possible for foreign authors to get U.S. copyrights for the first time.[16] Similarly, in 1972 it became possible for the first time to copyright musical recordings other than phonograph records—previously, such recordings were protected under other parts of the law. In 2000 6.8% of new copyrights were for sound recordings, so it is not so surprising that copyright registrations jumped up in 1972. In 1976, the term of copyright, which had been 28 years plus a renewal term of 28 years since 1909, was increased to the life of the author plus 50 years. In 1988 the United States eliminated the requirement of registering a copyright, so after that time, there is no reason to think of copyright registrations as a particularly good measure of the output of literary works. What all this means is that we should focus on the period between the major Copyright Acts of 1909 and 1972. Here we find that overall the literate population grew by 92%, while the number of copyright registrations grew by only 12%. Moreover, the literate population grew faster than the per capita copyright registrations in every decade, although in 1920–30 and 1960–70 the two growth rates are very similar. This is especially dramatic, because as we noted above, there was considerable technological change during the period, with entire new areas such as movies, recorded music, radio, and television opening up: by 2000 only 48% of new copyright registrations were for literary works—while in 1909 literary works accounted for the bulk of copyright registrations. Further, while the number of copyright registrations in the United States *overestimates* the share of the U.S. per capita labor dedicated to literary work, the size of the literate population grossly *underestimates* the size of the relevant market. The first is because a large number of foreign writers register their work in the United States, the second because the growth in the U.S. per capita income and, especially, the expansion of "American culture" around the world, greatly increased the potential market size.

The growth in per capita copyright is very low during the 70 years of stable legislation, suggesting this may be due to a binding constraint, rather than a gradual movement up the $M(\rho)$ curve. The labor constraint binding seems implausible in this context, there being, as far as we can tell, an essentially unlimited supply of mediocre want-to-be authors. Our model, however, implies an essentially unlimited demand for different books, while in fact the average reader may become satiated after reading, say, 50 books in a given year. We consider below the theory when there is satiation after a

[16] A brief history of U.S. copyright can be found at U.S. Copyright Office [2001a]. The 1972 change is described in U.S. Copyright Office [2001b].

certain number of ideas have been consumed. The implication is essentially the same as the labor constraint binding. Increasing protection beyond the minimum required to reach the satiation point enriches authors, while inefficiently leading to underconsumption of good books and overconsumption of mediocre ones.

4.2.6. *Analysis of Labor Demand: Patent Time Series.* We next turn to the demand for labor used to produce patentable ideas. One issue that arises is whether we should measure the scale of market λ by population or by GDP. Increases in per capita GDP increase the scale of the market, but they increase the opportunity cost of labor in the nonidea sector (working with existing ideas) by the same proportion, so have no impact on the effective scale of the market. On the other hand, increased productivity in the nonidea sector may also be reflected in increased productivity in the idea sector: double the per capita income may mean twice as many ideas. We will focus on population as a more conservative measure of λ in time series data, where per capita GDP is increasing. In the cross section we will examine both population and GDP as measures of scale of market.

Figure 8 is the patent analog of Figure 7, and is quite similar. Whether we measure patentable activity by patents awarded or patent applications, from 1890 to 1980 the growth rate of per capita patents exceeds the growth rate of population in only two decades, 1900–1910 and 1960–1970, and in both cases by only a trivial amount. In other decades, the growth rate of patents per capita is much lower than population growth, in some cases even negative. Overall, from 1890 to 1980 population grew at a rate of 1.4% per year, and per capita patents at 0.1% per year. Before 1890 patents per capita grew considerably faster than population, with a large drop in patents from 1860 to 1870 most likely because the reform of the patent law and patent office in 1861 made it considerably more difficult to get a patent. In the opposite direction, in the period after 1980 it became much easier to get and enforce a patent—the landmark event in this period being the formation of a special court to try patent cases in 1982. Needless to say, this court captured by patent attorneys has proven extremely hospitable to upholding questionable patents.

As the reader will recall, our empirical discussion is predicated on the assumption that $h(\omega) = 1$, a constant across ideas and time periods. While this is not so unreasonable for copyright data—a book taking roughly the same amount of labor input in 1909 as in 1972—labor input required to produce a patentable idea varies greatly both between different patents at a moment of time and over time. There is a vigorous debate over how to appropriately measure labor input into the production of patentable ideas. An alternative to measuring either patent applications or awards is to try to directly measure research and development expenditure. R&D expenditure, while in principle a better measure of input than patents, has a number of its own problems. First, the concept of R&D expenditure is itself fairly fuzzy

and available only for relatively recent years—the major source of data being a National Science Foundation survey conducted since 1953. The definition used by the NSF is "creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications." Firms and government agencies are surveyed and asked to report how much they spend on this activity.

The picture of R&D expenditure as measured by the NSF is radically different from that of the number of patents. Measured R&D expenditure excluding the federal government as a fraction of GDP roughly tripled between 1953 and 2002, and never grew by less than 14% in any decade. During this same time period population growth never exceeded 1.7% and real GDP growth averaged 3.2%. These data are discussed in Jones [2004]. Hence, if the NSF's measure of R&D expenditure by the private sector is an accurate measure of labor devoted to marketable ideas per capita, labor per capita devoted to innovation grew vastly more quickly than any measure of the scale of the market. Several comments are in order:

- Over time (even more so across countries) the tax treatment of activities regarded as R&D have changed. For example, the United States began a tax credit for R&D expenditure in 1981. It is easy to imagine that firms might wish to reclassify activities they had not previously regarded as R&D in response to tax incentives.
- The very notion of "R&D" as a special activity is recent. At one time these activities would have been called "entrepreneurship"—and it is far from clear if they would have satisfied the NSF definition of R&D. The amount of "formal" R&D conducted in laboratories by individuals with educational credentials has undoubtedly been increasing at the expense of "informal" R&D conducted by less educated individuals in homes, workshops, and shop floors. We wonder, for example, if the time spent by Jobs and Wozniak building the first Apple computer in their garage is included in NSF measured R&D for that year. If only the "formal" R&D is reported to the NSF this would lead to a substantial overstatement of the rate of growth of resources devoted to overall R&D. Note, however, that small individual inventors may be likely to patent small innovations so that they can sell them, while large firms may not patent small innovations that they plan to use internally. If organized research by large firms is supplanting that by small individual inventors, then this would have the effect that growth in "formal" R&D expenditure overstates the growth in actual R&D labor input, while the growth in patents per capita understates it.
- Dismissing the previous two "accounting" interpretations of the growth in R&D opens the door to a huge puzzle as to why when

resources per capita devoted to R&D grew at roughly 15% a year, patents per capita grew practically not at all.

- Total R&D, including the federally funded R&D, grew rapidly in the initial decade[17]—presumably as the Cold War got under way—and has remained relatively static since then, with declining federally funded R&D being replaced by private R&D. In 1958, for example, R&D per GDP is 0.23, while in the trough of 1994 it is 0.24 and peaks in 2002 at 0.278, which is slightly lower than its value in 1963, 0.284. If the propensity to patent from federally funded research were similar to that from privately funded research, there would be no patent puzzle. On the other hand, if the ratio of patents to R&D has remained relatively constant, then in the period prior to 1940, where there was little or no federal funding of research, our evidence above shows that per capita resources devoted to R&D were not growing as fast as the scale of the market. If we accept this view, then it is also possible that federally funded R&D was primarily displacing privately funded R&D—perhaps because the relevant labor constraint was binding—in which case private R&D was growing not because of the increased scale of the market, but because the federal government was reducing support for R&D.
- The growth of R&D expenditure will overestimate the growth in the demand for R&D labor if the cost of specialized labor grows faster than per capita GDP. This happens to be the case, and dramatically so. That the real wages of "stars" employed in all kinds of copyrightable productions have increased spectacularly faster than nominal GDP needs not be documented. The substantial raise of the wage premium for college graduates is also very well-known; between 1963 and 2002 the mean real wage of male college graduates has grown 64% while that of male high school graduates has grown about 31%. What is less well-known, but well documented (see Eckstein and Nagypal [2004]) is that the growth in the post-college premium was even more dramatic; during the same time interval the mean real wage of males with either a Master's or Ph.D. degree grew by more than 120%. Between 1963 and 2002, the period for which data are available, the ratio between nonfederal R&D expenditure and GDP goes from 0.095% to 0.20%, that is, it slightly more than doubles. During the same period, the ratio between the real wage of males with a post-college degree and the mean (median) real wage of males has increased 40% (65%). What this means is that, if one makes the reasonable assumption that the cost of labor employed in R&D grows like the wage of workers with a post-college degree, then the estimated share of actual workers employed in R&D over the total has

[17] Federally funded R&D per GDP increased by a factor of roughly 2.5 in that decade, and is 66% of total R&D expenditure in 1963.

grown roughly 50% during those 40 years, instead of tripling. While this number still cannot reject the hypothesis that the elasticity of total monopoly revenue is constant or decreasing, it is suddenly more consistent with the opposite one, i.e., that it is increasing.

4.2.7. *Analysis of Labor Demand: R&D Cross-Section.* Finally, we look at a cross section of countries. Here we run a simple regression with R&D as a fraction of GDP as the dependent variable and market size and the strength of IP protection as explanatory variables. There is an index of the strength of IPRs constructed by Walter Park.[18] We initially assume that the domestic market is what is significant. If ℓ represents per capita labor effort in the ideas sector and we assume constant elasticity, then $\log \ell = \vartheta \log \phi\lambda$. There is considerable evidence in the data the increases in per capita GDP positively influence the scale of the market. So we take $\lambda = \gamma^\alpha \pi$, where π is population and γ is per capita GDP. Ordinary least squares regression gives[19] $\vartheta = 0.20(0.03)$, $\alpha\vartheta = .56(0.038)$, meaning also that $\alpha = 2.9$, a remarkably large number.[20]

As can be seen, the elasticity with respect to λ is nowhere close to unity. However, this assumes that the relevant market for R&D is the domestic market. More generally, we would measure $\lambda = \lambda_{domestic} + \lambda_{world}$, where λ_{world} is the fraction of world GDP available as a market for domestic R&D. Since regressing log R&D on λ gives essentially the same result as regressing on $\lambda_{domestic}/\lambda_{average}$ and regressing on $\log(\lambda_{domestic}+\lambda_{world})$ gives essentially the same result as regressing on $\lambda_{domestic}/\lambda_{world}$, the regression coefficient should be multiplied by $\lambda_{world}/\lambda_{average}$. Thus, if the ratio of revenue earned on R&D in foreign markets to domestic markets is on the order of 5, it is possible that the elasticity of per capita R&D with respect to size of market is near unitary. However, this ratio is implausibly large. Since world GDP is on the order of $20,586 billion, the potential size of the world market is much larger than 5 times the GDP of even the largest country. However, this grossly overstates the relevant size of the world market. Exports are a fraction, not a multiple, of GDP. Consequently, a ratio of 5 would be possible only if R&D is much more intensive in export industries than domestic industries—by a factor of considerably more than 5. Using Lo's [2003] detailed data from Taiwan, in 1991 export-intensive industries spent about 1.8 times as much on R&D as domestic-oriented industries. Using microdata on

[18]We are grateful to Walter Park for providing us with his data. Details of the construction can be found in Park and Lippholdt [2003].

[19]Standard errors in parentheses.

[20]The underlying data include 34 countries in the period 1980–97. The R^2 for the regression is 0.65. There is no strong reason to think that ϕ is a linear function of the index of IP protection, which is computed as a score based on various legal criteria. However, if this measure of $\log \phi$ is included as a separate variable, the change in the estimates of ϑ, α is modest, while the t-statistic for the coefficient of $\log \phi$ is 1.711 and the corresponding increase in R^2 is only 0.008. The data on which these results are based can be found at http://www.dklevine.com/data.htm.

renewal rates to estimate the value of patents, Lanjouw, Pakes, and Putnam [1998] find the highest value of the "implicit subsidy from patenting abroad" at 35% for the United Kingdom and Germany, with most countries receiving 15–20% of income from a patent from rights held abroad. So the evidence hardly supports the idea that $\lambda_{world}/\lambda_{average}$ is on the order of 5.

4.2.8. *Conclusion on the Elasticity of Total Monopoly Revenue.* We examine a variety of data from different sources, ranging from book revenues to patent values estimated by renewal rates to R&D expenditures. We look at both cross-sectional and time-series data. Each individual analysis has many caveats—and this is clearly an area with a high rate of return to careful empirical work. In the case of copyright, we think that evidence in favor of increasing elasticity of total monopoly revenue is fairly decisive: all of the different sources of data say the same thing. In the case of patents, the evidence is less conclusive and far more subject to measurement problems, although we think it likely that the elasticity of total monopoly revenue is increasing in this case as well. In both cases, our best guess as to the functional form for $M(\rho)$ would be that it is approximately linear in the relevant range.

5. Extensions

5.1. **Variations on the Utility Function.** We assume that utility is linear in the output of the idea sector and in labor. We can consider more generally the functional form

$$U\left(g(\lambda)\int_{\underline{\rho}}^{\infty}[\phi\overline{\nu}^{M}+(1-\phi)\overline{\nu}^{C}]\mu(\rho)d\rho\right)$$
$$+V\left(L-(1/\lambda)\int_{\underline{\rho}}^{\infty}\mu(\rho)d\rho\right).$$

In our examination of copyright time series we found reason to suspect that the lack of growth in per capita copyright is due to a limit in the per capita demand for books. The empirically relevant model is one in which $U(\cdot)$ displays, after a point, sharply decreasing marginal utility. To see how this works, consider the limit case in which $U = u$ for $u \leq \overline{u}$, and then $U(u) = \overline{u}$ for $u > \overline{u}$; continue to assume V is linear all thorough its domain. As long as λ is small, u will be small enough that satiation cannot occur, and the price of output in the idea sector is $P = 1$. Once λ grows, the satiation constraint eventually binds. Equilibrium requires that exactly $\overline{u}$ be produced in the idea sector, so the price in the idea sector falls to $P < 1$ to discourage labor from flowing into that sector. This is the mirror image of the labor constraint binding.

More general nonlinearities in U and V have a similar effect. As the size of the market grows and more ideas are produced, the price in the idea sector U' declines for fixed ϕ, and the marginal utility of labor in the other sector

V' increases as labor moves to the idea sector and the relative price of labor goes up. In general it is still best to exploit the opportunity offered by an increase in the size of the market by reducing ϕ, rather than by allowing the relative price of skilled labor to rise.

5.2. **Return Nonneutrality.** So far we have focused on return neutrality. Note that, as a function of private return under monopoly, social value under monopoly is $\overline{\nu}^M(\rho)\rho$ while social value under competition is $\overline{\nu}^C(\rho)\rho$ so that, for given ϕ, expected social value is $\rho[\phi\overline{\nu}^M(\rho)+(1-\phi)\overline{\nu}^C(\rho)]$. The two polar cases in which private return and social values are, respectively, positively and negatively related are worth considering.

If goods with lower private return also have lower social value, in the sense that $D\overline{\nu}^M(\rho) > 0$ or $D\overline{\nu}^C(\rho) > 0$, or both, common sense and simple calculations show that this is a further reason for the length of protection to decline with the scale of the market. On the other hand, if $D\overline{\nu}^M(\rho) < 0$ or $D\overline{\nu}^C(\rho) < 0$, or both, then this weakens the connection between the scale of the market and the declining optimal protection. In particular, in this case, it becomes possible to have the optimal degree of IP protection increasing with market scale, even when the elasticity of total monopoly revenue is increasing, so that the labor constraint never binds.

On the other hand, while $D\overline{\nu}^M(\rho) < 0$, $D\overline{\nu}^C(\rho) < 0$ may seem to reinforce the case for increasing IP protection; in fact, it weakens it. That is, $D\overline{\nu}^M(\rho) < 0$, $D\overline{\nu}^C(\rho) < 0$ means that private return is poorly correlated with public benefit. In the extreme case, there may actually be a negative correlation between private and public benefit. In this case, the private sector produces the ideas of least social merit first. When the market is small, IP protection results only in the production of ideas of little social value, so scarcely makes sense. The argument for strengthening protection as the scale of the market increases is that the increased scale of the market eventually leads the private sector to produce ideas that do have some significant social value, and at this point, we can try to compensate for the weakness of private incentives by increasing the level of protection. While this is formally correct, it clearly is a lopsided argument when it comes to designing welfare-improving policies. It takes as given the policy instrument and patents, even if the latter is the least adequate to maximize social welfare. If it were really the case in practice that privately valuable innovations have little or no social value, and vice versa, then a form of government intervention other than IP would be sensible, such as publicly sponsored research projects, or auctioning of production rights, or subsidies for innovators producing the socially valuable ideas, for example. Patents, certainly not.

Alternatives to Government Grants of Monopoly. The latter remarks, that when private and social values of new ideas are not aligned, IP protection is the least appropriate policy instrument, suggest we should, albeit briefly,

consider how alternative forms of government intervention fare in our environment. An obvious alternative to having the government award private monopolies is to have the government award prizes for innovation. This can be financed in much the same way that private monopolies raise money—by imposing a sales tax on sales of new goods. Unlike the award of a private monopoly, the tax rate does not need to be set so high as to give the monopoly revenue, and Gilbert and Shapiro [1990] show in effect that having a low tax throughout the life of the good is preferable to having a high tax (monopoly) for part of the life of a good and low tax (after the copyright/patent expires) for the remaining life of the good. Hence, such a system of taxes is intrinsically less distortionary than awarding private monopolies. Insofar as the prize money is simply paid back to the innovator, this is essentially the same as a system of mandatory licensing, in which the holder of the private monopoly is required to sell at government mandated prices. Systems of mandatory licensing are widely used—in copyright, for example, such things as radio play of music and photocopying of copyrighted materials are covered by mandatory licensing provisions. In the case of patent, mandatory licensing was widely used in Taiwan until the Taiwanese were forced to reform their patent system by the United States. So this kind of mandatory licensing represents, as we might expect, the efficiency improvement from replacing an unregulated monopoly with a regulated monopoly.

However, there is little reason that the proceeds of taxes on new goods should be paid back to the innovator. From an efficiency perspective, it is better that the proceeds be used to defray the costs of producing innovations of high social value. This has several advantages over an intellectual property system. First, to minimize the monopoly/tax distortion, the minimum necessary to get innovation should be paid. In particular, it is best to pay $h(\omega)$, the indivisibility, to the innovator rather than the full social value. The intellectual property system makes little use of social knowledge of $h(\omega)$; with the exception of the nonobviousness requirement (now largely defunct) of patent law, patents and copyright base reward on social value rather than social cost. Second, as we noted above, if it is indeed the case that social value is poorly correlated with private value (the strongest case for increasing intellectual property protection as the scale of the market increases) a system of rewards based on other information about social value is likely to lead to a much better mix of innovations being produced. It is important to note that, like mandatory licensing, systems of public (and private) prizes have been widely used and are of demonstrated practicality.

The issue, also in the context of our model, boils down to the public knowledge of the true social cost of introducing a new idea. When the latter is known, a public subsidy to innovators equal to the amount $h(\omega)$, financed by a consumption tax and followed by unconstrained competition, is easily shown to provide the least distortionary mechanism. When the information about the true cost $h(\omega)$ of innovating is private the problem appears less straightforward and worthy of further investigation.

5.3. **Positive Marginal Cost of Distributing Ideas.** So far we have assumed that there is no marginal cost of reproducing and distributing ideas or goods that use ideas. There are several possible cases, depending on which inputs are needed to do this. One possibility is that the same labor used to create ideas is used to reproduce them. This case is rather complex, because it introduces a third margin into the choice of ϕ—the monopoly for inframarginal ideas, the marginal ideas, and the amount of labor used to reproduce existing ideas.

In practice the type of labor used to reproduce ideas is probably not a terribly good substitute for the labor used to produce the ideas themselves. If we introduce an additional factor of production—unskilled labor, call it—and assume that this is used to reproduce ideas, provided this factor is in plentiful supply so that marginal cost is constant, little change is needed in our analysis. In particular, instead of examining $v(z,\omega)$, we should examine per capita utility net of the cost of reproducing the idea: $v(z,\omega) - mc(\omega)z$. This may have an impact on whether quality is neutral, since it may be neutral for $v(z,\omega)$, but not for $v(z,\omega) - mc(\omega)z$, but, for example, with quadratic utility (linear demand) we have neutrality in both cases. In general, we would expect this modification to make it more likely that private return and social value of an idea move together. This is because one intuitive reason social value increases with private return is that the marginal cost of producing additional copies of an idea increases more slowly than utility from the additional copies. As we have seen this reinforces our main prediction, that IP protection should decrease when market size increases.

If the unskilled labor constraint binds before the skilled labor constraint does—something we think is unlikely, but since the demand for unskilled labor in this model grows much more rapidly than the demand for skilled labor as λ increases, something we recognize is a possibility—then the marginal cost of producing copies increases as output of ideas increases. However, raising ϕ also does not generally result in an increase of the production of new ideas, but rather raises the cost of producing copies of old ones, offsetting the gain to the innovator from higher ϕ. Here the additional monopoly profits accrue to the scarce factor—unskilled labor. But the effect is not different than if the constraint for skilled labor was binding—there is no efficiency gain in creating a socially inefficient monopoly merely in order to increase the income of one group at the expense of another.

5.4. **Consequences of Competitive Rents.** As argued in Boldrin and Levine [1999, 2002, 2004a,b] it is by no means true that in the absence of any IP protection profits for innovators are negligible or even zero. At each moment of time, and especially shortly after innovation took place, a capacity constraint is present that will give rise to nondistortionary competitive rents. Most likely, there are also first-mover advantages, such as those documented by Tofuno [1989] in the market for financial securities.

We can model this by assuming a capacity constraint $\overline{z}(\omega)$ on per capita production after the IP protection expires. In this case, assuming the capacity constraint is not binding during the IP protection period, the innovator's total revenue from an idea ω is

$$\phi\lambda z^M(\omega)v_z(z^M(\omega),\omega) + (1-\phi)\lambda\overline{z}(\omega)v_z(\overline{z}(\omega),\omega).$$

Assuming the labor constraint does not bind, hence $w = 1$, dividing through by $h(\omega)$, we can write the condition for good ω to be produced as

$$\phi\lambda\rho(\omega) + (1-\phi)\lambda\rho^C(\omega) \geq 1,$$

where $\rho^C(\omega) = \overline{z}(\omega)v_z(\overline{z}(\omega),\omega)/h(\omega)$ is the competitive rent per unit of indivisibility cost. Let us use the simplifying assumption that competitive rent is proportional to monopoly revenue, per unit of indivisibility cost; that is, $\rho^C(\omega) = \vartheta\rho(\omega)$, with $0 < \vartheta < 1$. Note that this is stronger than our earlier neutrality assumptions, which had to hold only in expected value. Then we can again write social welfare entirely in terms of ρ, and the only modification of our earlier expression for social welfare is that $\underline{\rho} = [(\vartheta + \phi(1-\vartheta))\lambda]^{-1}$. The corresponding NOC is

$$\begin{aligned} NOC(\lambda,\phi) = (1-\vartheta)&\left[(1/(\vartheta+\phi(1-\vartheta)))\left\{\phi\overline{\nu}^M + (1-\phi)\overline{\nu}^C\right\} - 1\right] \\ &\times \Upsilon(1/(\vartheta+\phi(1-\vartheta))\lambda) \\ &- (\overline{\nu}^C - \overline{\nu}^M), \end{aligned}$$

from which we see that it continues to be true that if

$$\Upsilon'(1/(\vartheta+\phi(1-\vartheta))\lambda) > 0,$$

then $\hat{\phi}(\lambda)$ is unique and nondecreasing. In this case, we can also check that $NOC_\vartheta < 0$, so that higher competitive rents lead to a reduction in the optimal level of protection.

Finally, consider that the NOC at $\phi = 0$ is

$$(1-\vartheta)\left[(1/\vartheta)\overline{\nu}^C - 1\right]\Upsilon(1/\vartheta\lambda) - (\overline{\nu}^C - \overline{\nu}^M),$$

and as $\lambda \to \infty$ this approaches $-(\overline{\nu}^C - \overline{\nu}^M) < 0$. Hence, when competitive rents are present, the optimal level of protection should be set to zero at a finite market size, not just asymptotically as in the extreme case of no competitive rents.

5.5. **Consequences of Rent-Seeking.** Suppose that the size of the indivisibility does not vary systematically with private return. Then ideas with high returns also have high absolute levels of profit associated with them. Suppose it is possible to purchase "extensions" of protection, in the form of a $\Delta\phi$, from the government sector at a cost. Then it is owners of ideas with high ρ that have the greatest incentive to do so, as they can "leverage" the $\Delta\phi$ more than anyone else. This means that the marginal firms, who from a social point of view are the reason for IP protection, do not get much say over the length of protection. In the extreme case the marginal firms get no protection, so the set of ideas produced is the same as without IP, and IP

serves only to introduce a monopoly distortion. Rather remarkably, Landes and Posner [2003] recommend embodying such a scheme in law.

6. International Trade and Harmonization

We now turn to the issue of IP protection in the world economy. Since it is the empirically relevant case, we assume throughout this section that the elasticity of total monopoly revenue is increasing. Our goal is to examine whether optimal trade harmonization, meaning that all countries must set the same level of IP, results in countries increasing or decreasing their level of IP.[21] It is tempting to view this as a typical tariff-like free-riding problem: countries try to free-ride off of each other's innovation, and harmonization enables them to agree on a more efficient higher mutual level of protection. But, because of the scale-of-market effect, this need not be the case. We find that—in the empirically relevant case where only a small share of large country patent revenue flows to small countries—harmonization demands that the large countries lower their level of protection. Basically there are two effects: one is that there is a tendency to underprotect because some royalties are lost to overseas innovators. The second is the scale-of-market effect, which works in the opposite direction.[22]

Consider first the simple case in which there are several countries and no trade is possible. In this case opening the economies to trade in goods and ideas results only in a scale-of-market effect, and given our maintained assumption that the elasticity of total monopoly revenue is increasing, by our previous analysis the welfare optimum for the set of countries as a whole is to reduce protection. The more demanding case is to consider a situation in which there is already trade, but countries engage in non-cooperative individually optimal IP policies before harmonization sets in.

We assume that there are I countries and that each country i has a fixed fraction θ_i of world demand and labor and a fixed fraction ϵ_i of world ideas. The total size of the world economy is still λ. We focus on the case in which countries may not discriminate against foreign inventors. While *de facto* violated in some occasions, this reflects current legal practices around the world, and it allows us to focus on the specific role of IP protection. We let ϕ_i denote the level of IP protection in country i. Our base assumptions are that there is complete and costless free trade of goods, that the labor

[21] Grossman and Lai [2004] argue that harmonization necessarily results in all countries increasing their level of IP. Due to an algebraic error, their argument is invalid except in the constant elasticity case, as we explain below.

[22] Grossman and Lai [2004] correctly identify the first effect, but due to an algebraic error, miss the second. In footnote 28 they compare the first-order condition for a harmonized welfare maximum to a best response in which the foreign country does not protect. In this comparison they treat their parameter γ, which is the same as our Υ, as having the same value in both equations. This is true only in the constant elasticity case, meaning for their production function approach the production function must be Cobb-Douglas and in our distributional approach the distribution must be Pareto.

constraint does not bind, and that the elasticity of total monopoly revenue is increasing.

From an inventor's perspective, what is relevant is the effective (weighted by market shares) total protection received worldwide. This is simply $\phi = \sum_i \phi_i \theta_i$, hence $\underline{\rho} = 1/\phi\lambda$ determines the marginal invention worldwide. Each country is supposed to pick ϕ_i to maximize its own welfare,

$$\theta_i \lambda g(\lambda) \int_{\underline{\rho}}^{\infty} [\phi_i \overline{\nu}^M \rho + (1-\phi_i)\overline{\nu}^C \rho] \mu(\rho) d\rho + L \qquad +$$
$$+\epsilon_i g(\lambda) \int_{\underline{\rho}}^{\infty} [\phi\lambda\rho - 1]\mu(\rho)d\rho -$$
$$-\phi_i \theta_i \lambda g(\lambda) \int_{\underline{\rho}}^{\infty} \rho\mu(\rho)d\rho.$$

The first component is the total utility that agents in country i receive from their consumption of goods and leisure. The second is the profits accruing to the monopolists located in that country, which is the difference between the revenue from worldwide sales and the cost of labor used to innovate. The third is the total expenditure of consumers in country i for their purchases of goods. To get per capita welfare, we normalize this by the country population $\theta_i\lambda$. In the case of a closed economy, $\epsilon_i = 1$ and $\phi_i = \phi$, so the profit of the monopolists minus the consumers' total expenditure is simply equal to the cost of production, getting us back to the single country social welfare function above.

The optimum of an individual country is calculated by differentiating the social welfare function to get

$$NOC(\phi_i, \phi) =$$
$$\frac{\theta_i}{\phi}\left[\phi_i(\overline{\nu}^M - 1) + (1-\phi_i)\overline{\nu}^C\right] \Upsilon(1/\phi\lambda) - (\overline{\nu}^C - \overline{\nu}^M + 1 - \epsilon_i).$$

Because the elasticity of total monopoly revenue is assumed increasing, this is strictly concave in ϕ_i and continuous as a function of (ϕ_i, ϕ), so the IP protection game has a pure strategy Nash equilibrium, characterized by the first-order conditions $NOC(\phi_i, \phi) = 0$.

The Symmetric Case. Consider first a symmetric equilibrium of a symmetric model in which $\theta_i = \epsilon_i = 1/I$. In equilibrium we must have $\phi = \phi_i$, which gives

$$NOC(\phi, \phi) =$$
$$\left[\phi\left(\overline{\nu}^M - 1\right) + (1-\phi)\overline{\nu}^C\right] \frac{\Upsilon(1/\phi\lambda)}{I\phi} - (\overline{\nu}^C - \overline{\nu}^M + 1 - (1/I)) = 0.$$

Holding constant the total market size λ, let ϕ^1 be the solution to the single country problem from Section 3, and $\phi^I$the symmetric solution to

the equation above. Because the first term is positive and the second negative ($I > 1$), and because the second term becomes more negative as I increases, we have that $NOC(\phi^1, \phi^1) < 0$, implying, since $NOC_{I\phi} < 0$ under the elasticity condition, that $\phi^I < \phi^1$. More generally, ϕ^I is decreasing in I and as the number of countries increases, the symmetric Nash equilibrium converges to the case of no IP, which is suboptimal in this setting. The intuition behind this result is simple and ordinary: by decreasing ϕ^I a country loses because it creates fewer new goods and gains because it consumes at the competitive level the goods created by the remaining $I-1$ countries. As I increases the second margin strictly dominates the first.

The NOC for the single country problem coincides with the social optimum for a global economy. Moreover, if each country is constrained to set the same level of protection as all others, for example through a legal mechanism such as the WTO, they would all agree to choose the social optimum ϕ^1. This is the standard harmonization result: in the unconstrained protection game countries underprotect due to the public goods nature of IP protection, and a WTO-like mechanism that forces harmonization leads them to the second best.

North versus South. Unfortunately, this analysis has little normative relevance to policy analysis. Current extensions of IP are not between countries of equal size with currently equal levels of IP. Rather, extension of IP protection is taking place between two very heterogeneous groups of countries. The first, consisting of North America, Europe, and Japan, has a relatively large θ_i and an even much larger ϵ_i, and has been harmonized for around a century on a high level of IP protection. The second group consists of developing countries with little or no IP protection. For purposes of calibration, it is convenient to focus on the most significant of these countries: Brazil, China, India, Mexico, and Russia. The relevant facts about GDP in these countries and their share of U.S. patents from Hall [2001] are shown in Table 5.

Table 5

	GDP Trillion USD	%World GDP (θ_i)	%US Patents (ε_i)
Brazil	1.13	2.59	0.07
China	4.50	10.32	0.86
India	2.20	5.05	0.11
Mexico	0.92	2.10	0.05
Russia	1.12	2.57	0.14
Total	9.87	22.63	1.23
World	43.60	100	100

For the purposes of our numerical exercise it makes sense to assume the "North" controls about $\theta_1 = .75$ of world GDP and $\epsilon_1 = .987$ of world ideas. We will also focus on the case in which demand is linear, so $\overline{\nu}^M = 3/2, \overline{\nu}^C =$

2, and in which total monopoly revenue $M(\rho)$ is also linear—as this seems best supported by the data.

Why There Is No IP in the South. First we show as a matter of theory that, regardless of I, the equilibrium level of aggregate protection, ϕ, is bounded away from zero. There is one large country with shares θ_1, ϵ_1 and $I-1$ small countries with shares $\theta_i = (1-\theta_1)/(I-1) < \theta_1$, $\epsilon_i = (1-\epsilon_1)/(I-1) < \epsilon_1$. The NOC for the large country is

$$NOC_1(\phi_1, \phi) = \frac{\theta_1}{\phi}\left[\phi_1\left(\overline{\nu}^M - 1\right) + (1-\phi_1)\overline{\nu}^C\right]\Upsilon(1/\phi\lambda) - (\overline{\nu}^C - \overline{\nu}^M + 1 - \epsilon_1) = 0.$$

Observe that $\phi \geq \theta_1\phi_1$ and recall that $NOC_1(\phi_1, \phi)$ is decreasing in ϕ. Hence, $NOC(\phi_1, \theta_1\phi_1) \geq 0$. Since this latter expression is also decreasing in ϕ_1, a solution to $NOC(\widetilde{\phi}_1, \theta_1\widetilde{\phi}_1) = 0$ must satisfy $\phi_1 \geq \widetilde{\phi}_1 > 0$. This in turn implies that in equilibrium $\phi \geq \theta_1\widetilde{\phi}_1 > 0$. This shows that ϕ is bounded away from zero independent of k because the large country will never impose a negligible amount of protection.

We now turn to the NOC for the small countries. At $\phi_i = 0$ this is

$$\begin{aligned} NOC_i(0, \phi) &= \frac{(1-\theta_1)}{(I-1)}\left[\frac{1}{\phi}\overline{\nu}^C - 1\right]\Upsilon(1/\phi\lambda) - (\overline{\nu}^C - \overline{\nu}^M + 1 - \frac{\epsilon_1}{I-1}) \\ &\leq \frac{(1-\theta_1)}{(I-1)}\left[\frac{1}{\widetilde{\phi}_1}\overline{\nu}^C - 1\right]\Upsilon(1/\theta_1\widetilde{\phi}_1\lambda) - (\overline{\nu}^C - \overline{\nu}^M + 1 - \frac{\epsilon_1}{I-1}), \end{aligned}$$

which is strictly negative for I larger than

$$I^* = \frac{(\overline{\nu}^C - \overline{\nu}^M + 1)}{(1-\theta_1)\left[(1/\theta_1\widetilde{\phi}_1)\overline{\nu}^C - 1\right]\Upsilon(1/\theta_1\widetilde{\phi}_1\lambda) + \epsilon_1} + 1.$$

Since there is always a unique solution to $NOC_i(\phi_i, \phi) = 0$, for $I > I^*$ it occurs at $\phi_i = 0$.

Turning from the theory to the calibration, with linear demand, the NOC is

$$NOC(\phi_i, \phi) = \frac{\theta_i}{\phi}\left[(1/2)\phi_i + (1-\phi_i)2\right]\Upsilon(1/\phi\lambda) - ((3/2) - \epsilon_i) = 0.$$

From our earlier analysis, we know that a plausible range for Υ is 0.15–0.40, while for plausible interest rates the current U.S. patent term corresponds to a ϕ_{US} in the range 0.3–0.6. Assuming $\epsilon_i = \theta_i$ and plugging in $\phi_i = 0$ for a small country, we see that it is indeed optimal for a small country to set $\phi_i = 0$ if

$$\theta_i \leq \frac{3}{(4\Upsilon/\phi) + 2}.$$

If we assume that the part of the world setting ϕ_1 at least 0.3 consists of at least two-thirds of the world economy—which is true of the G7 alone—then ϕ is at least 0.2, while $\Upsilon \leq 0.4$. This gives a lower bound for the right-hand side of 0.3, so any country with a smaller fraction of world GDP than this

should not protect at all. Since none of the "Southern" countries control near this fraction of world GDP, this lower bound is abundantly satisfied. Note that if, contrary to our assumption, ϵ_i is smaller than θ_i, as it is, then there is even less incentive for the South to choose a positive level of IP protection in the current circumstances.

Should the North Increase or Decrease IP under Harmonization? Consider first the case in which $\epsilon_1 = 1$ and $\theta_1 < 1$, that is, all ideas are produced in the large country. In this case the solution for the large country is the solution to the social optimum problem, which ignores supply from the rest of the world and chooses ϕ_1 to be optimal for a population of $\theta_1\lambda < \lambda$. By the usual scale of market effect, that means the equilibrium solution for ϕ_1 is larger than the value that maximizes world social welfare, that is, the solution to the social optimum problem with population λ.

When some ideas are produced in the smaller country, that is, $\epsilon_1 < 1$, this effect is weakened. This is because of the "profit-stealing" effect mentioned above. When the small countries are not protecting at all, the large country has an incentive to set a lower level of protection than if it were the only country in the world. The reason is that if it were the only country in the world, it would retain all the royalties from IP. However, if the smaller countries also produce ideas, then part of the increased royalty payments is lost to overseas monopolists. Since in this case the large country sets a lower level of IP than when $\epsilon_1 = 1$, the profit-stealing effect tends to offset the scale-of-market effect from increasing IP through harmonization in the small countries. In the extreme case of constant elasticity of total monopoly revenue—the Cobb-Douglas/Pareto case—the scale-of-market effect vanishes and only the profit-stealing effect is left, so that it is unambiguous that the large country should also increase IP with harmonization. This is the case studied by Grossman and Lai [2004].

In our calibration, the NOC for a single aggregate large country is

$$[(1/\phi_1)\{(1/2)\phi_1 + (1-\phi_1)2\}]\,\Upsilon(1/\theta_1\phi_1\lambda) = ((3/2) - \epsilon_1),$$

while the NOC for the harmonized welfare maximum is

$$[(1/\phi)\{(1/2)\phi + (1-\phi)2\}]\,\Upsilon(1/\phi\lambda) = 1/2.$$

We take $\theta_1 = .75$ throughout. Then the RHS of the single aggregate large country NOC is $3-2\epsilon_1$ times the RHS for the harmonized welfare maximum. On the other hand, if world protection under harmonization is the same as the optimum for the single aggregate large country, that is, the optimal $\phi = \phi_1$, then the LHS of the single aggregate large country NOC is the elasticity of Υ times $1/\theta_1$. In the case in which $M(\rho)$ is linear, the elasticity $\rho\Upsilon'/\Upsilon = 1 + \Upsilon$, so the LHS is $(1+\Upsilon)/\theta_1$. By the second-order condition, it follows that it is optimal for the single large country to reduce protection to reach the harmonized optimum if and only if $(1+\Upsilon)/\theta_1 > 3 - 2\epsilon_1$.

Suppose first that the "South" is just as effective at producing ideas as the "North" so that $\epsilon_i = \theta_i = .75$. In this case we see that the large country

should reduce protection upon harmonization if $\Upsilon \geq .33$, which is near, but still below, the upper range of our estimates of Υ.

However, the assumption that $\epsilon_i = \theta_i$ is not nearly true. As Table 5 showed, the "South" controls on the order of 25% of world GDP, but generates well less than 2% of all U.S. patents. That is, the "profit-stealing" effect is quite small, hence there is little reason for the "North" in the pre-harmonization equilibrium to decrease its protection in order to decrease the trivial revenue from patents it loses to the "South." In our calibration, we have taken $\epsilon_1 = 0.987$. With this calibration $1/\theta_1 > 3 - 2\epsilon_1$ so that, upon harmonization, the"North" should reduce protection regardless of the elasticity Υ.

To estimate what the actual reduction of the Northern IP should be upon harmonization in the calibrated version of our model, we compare the solution to the NOC for the single aggregate large country with that for the harmonized welfare maximum, i.e., the two values of ϕ that solve our two NOC above. Let ϕ_j be the optimum for the jth of the two NOCS, RHS_j be the RHS of the jth NOC, and Υ_j the elasticity of the jth NOC. Then we can solve the jth NOC to get

$$\phi_j = \frac{4}{3 + 2RHS_j/\Upsilon_j}.$$

Dividing the two solutions and using the approximation that $\Upsilon_1/\Upsilon \approx (1 + \Upsilon_1)(\phi/\phi_1\theta_1)$

$$\begin{aligned} \frac{\phi}{\phi_1} &= \frac{3 + (3 - 2\epsilon_1)/\Upsilon_1}{3 + 1/\Upsilon} \\ &= \frac{3\Upsilon_1 + (3 - 2\epsilon_1)}{3\Upsilon_1 + (1 + \Upsilon_1)(\phi/\phi_1\theta_1)}. \end{aligned}$$

The resulting quadratic solves as

$$\frac{\phi}{\phi_1} = \frac{-3\Upsilon_1 + \sqrt{9\left(\Upsilon_1\right)^2 + 4\left(3\Upsilon_1 + (3 - 2\epsilon_1)\right)\left(1 + \Upsilon_1\right)/\theta_1}}{2\left(1 + \Upsilon_1\right)/\theta_1}.$$

For Υ in the range 0.15–0.40 this implies ϕ/ϕ_1 in the range 0.817 to 0.845. In other words as part of TRIPs it would make sense to have about a 15–20% reduction in length of patent terms, from, say, 17 years to 14 years.

Copyright. The astute reader will have noticed we have examined only the harmonization of patents, and not that of copyrights. Current copyright protection is effectively infinite while elasticities appear to be extremely low. As we observed, this means that copyright protection is not consistent with welfare maximization by the "North." What this means is that the social planner in the North is not setting current copyright levels to maximize social welfare but, we conjecture, to maximize the rents accruing to the interest groups that are covered by the copyright laws. Hence, while this means that harmonization from *optimal* levels of protection might result in an increase

in protection, current levels so greatly exceed the optimum that either with or without harmonization they need to be drastically cut, and countries not currently providing copyright protection would be foolish to agree to provide such protection as part of a negotiated international agreement.

Other Considerations. Our base assumption is that countries cannot increase the level of domestic innovation by changing their IP laws, as their share of total innovation is fixed at ϵ_i. Insofar as countries can increase their share of worldwide innovative production by changing their national level of IP protection, they benefit from the fact that they increase their share of the total monopoly profits from innovation. In fact there is some evidence that favorable IP treatment can attract innovation. There are several reasons for this. First, favorable IP legislation may be a signal of favorable treatment of innovators in general (for example, as in Ireland, through a generous tax treatment of profits from FDI). Second, although legal discrimination against foreign inventors is forbidden in principle, there may be a variety of informal reasons why it is advantageous to be a domestic innovator to take advantage of strong local IP protection. Finally, the distribution of innovation across countries can be driven by the explicit rent-seeking behavior of innovators, who may choose to reward countries that provide favorable IP protection with increased revenue from domestic innovation.

Insofar as increasing IP protection lures innovation, a second type of equilibrium distortion arises. Rather than underprotecting in an effort to free ride off of innovation in other countries, the incentive is to overprotect to try to get a disproportionate share of IP revenue.

7. Conclusion

7.1. **Relation with Previous Literature.** Besides the work of Grossman and Lai [2002, 2004] there is a wide partial equilibrium literature on the optimal length of patent protection. This literature, stemming from the paper of Gilbert and Shapiro [1990], examines the trade-off between patent length and breadth for a single innovation. Gilbert and Shapiro give assumptions under which optimal length is infinite, while Gallini [1992] shows that with a more realistic model of the "breadth" of protection, this result may be reversed. This literature does not examine the broader question of optimal policy that covers many different ideas, and takes as given that policy can easily determine "breadth" as well as "length." We think that "breadth" is much more difficult to legislate than "length," and because it is less visible, more subject to rent-seeking, regardless of legislative intent. In our model, unlike this literature, we effectively take the "breadth" of protection as exogenous and focus on length. Insofar as "breadth" as well as "length" can be legislated, our parameter ϕ can be regarded as kind of a summary of length and breadth combined. Hence, it would be good public policy to reduce breadth as the scale of the market increases as well as length. In particular, it might be a good idea to introduce the "independent invention" defense as

suggested by Maurer and Scotchmer [2002], or to eliminate product patents in favor of process patents only. However, it is clearly more practical to tie a time limit to the growth of the economy than a particular scope of coverage.

A number of reasons, other than the invariance of the optimal length of patents to the size of the market, have been invoked to justify policies asking for an increase of the degree of IP protection as a consequence/condition of trade liberalization. First, as Diwan and Rodrik [1991] argue, northern and southern countries generally have different technology needs and, without the southern protection of IPRs, northern countries would not develop technologies largely needed by the South. This seems rather at odds with both basic principles and historical facts. Developing countries have developed, almost always and invariably, by adopting technologies that had been developed by the advanced ones. Indeed, the very same notion of "convergence" would make no sense if this were not the case. Examples of countries that have developed late by inventing technologies substantially different from those invented by the already developed ones are conspicuously absent. China and India are currently achieving unusually high growth rates by adopting *exactly* the same technologies adopted earlier by developed countries; this is currently called "outsourcing," and it means just that. So much for history and facts. But even in an abstract and purely theoretical sense, the Diwan and Rodrik story can make sense if and only if one argues that poor countries have natural and unmodifiable resources and skills that are completely different and orthogonal to those that richer countries had when they were at the same stage of development. Further, one needs to also assume, or show, that the technologies adopted by advanced countries can neither be learned nor transported to the poorer ones, or that doing so is more costly than developing completely new, and as of now yet unknown, technologies. Both claims sound implausible. All historical evidence, including the development of the United States in the nineteenth century, shows that "convergence" or "catching-up" takes place only when the followers imitate the technologies of the leaders. Hence, facilitating technological imitation (call it "pirating" if you like) is the *key* to economic development for newcomers. No imitation, no (development) party.

A second line of thought argues that northern firms may react to the lack of IPRs in the South by making their technologies more difficult to imitate, which can result in less efficient research technologies and fewer northern innovations (Taylor, [1993, 1994]; Yang and Maskus [2001]). This is possible in theory; facts seem to contradict it, though. In theory, the opposite effect is also possible: the more the South imitates the North, the more the Northern firms innovate to preserve the value of their immobile factors. In this case, lack of IP protection in the South fosters higher rates of innovation in the North. Historically, IP protection in the underdeveloped countries was weaker in the nineteenth century than now, and this did not seem to lead firms in Europe to invest heavily in secrecy or innovate less. If secrecy was so high when, say, the United Kingdom was the leading country

and most of economically backward Europe had no IP protection laws, how did those European countries ever manage to imitate and adopt the British technologies? The same, rhetorical, question could be asked for Japan, South Korea, and a long list of countries that developed during the last 50 years.

Third, some commentators on intellectual property, such as Landes and Posner [2003], have become confused on the point of optimal length of intellectual property protection, arguing for a system of perpetual copyright renewal that would have little effect in increasing the incentive to innovation, while perpetuating the monopoly losses on inframarginal productions. There is no sense in which such informal arguments, often developed by confusing "pecuniary" with "real" externalities, can be derived from a general equilibrium theory of optimal IP protection.

Finally, it should be noted that the theory proposed here does not say that the South should have systematically less IP protection than the North but, instead, that both should have a common IP protection at a level substantially lower than it is currently in the North, e.g., the United States and the European Union.

7.2. **Shortcomings and Future Research.** The most important missing aspect of our analysis is the dynamic feature that ideas build on other ideas. As pointed out in Scotchmer [1991] and Boldrin and Levine [1999, 2004a, b], ideas that use other ideas as inputs greatly weaken the case for IP because the latter, while it encourages innovations by improving the return to the first inventor, discourages further innovations through raising their cost. In this sense, there is no reason to think that adding true dynamic features to the model is likely to make IP more socially desirable. In fact, when the complexity of innovations increases because new ones need to use more and more old ideas as inputs, the presence of widespread IP naturally determines a holdup problem where even one residual monopolist can prevent new ideas from being implemented. (See Boldrin and Levine [2004b] for a simple formal version of this argument.)

As we observed, neither that the number of ideas can increase with size at different rates for different characteristics nor that the indivisibility varies with the size of the economy is a possibility considered here. These are valuable, but theoretically demanding, extensions for various reasons. The first extension is valuable because new ideas are not all equally useful; hence, the fact that market size can make ideas with certain characteristics more abundant than others is relevant for welfare. The relevance comes from the fact that ideas with high private returns need not be ideas with high social values, and vice versa; an optimal innovation policy should be designed by taking this effect into account. This is a hard problem to tackle on purely theoretical grounds, and, unfortunately, we are aware of no empirical research measuring the relationship between private return and social value of new ideas.

The second extension, allowing for variations in the indivisibility $h(\omega)$ that are due to variations in market size, is also interesting. In theory, one can argue that a larger economy implies more competition to develop ideas potentially already available, and few additional potential ideas; that is, $h(\omega, \lambda)$ increases with λ while $g(\lambda)$ is practically flat. The opposite case is, clearly, also a theoretical possibility. Which of the two is empirically relevant would be important to provide guidance to theoretical research. To discriminate between the two polar hypotheses one needs to be able to measure the portion of $h(\omega)$ that is wasted when one does not win the innovation race, versus the complementary portion that may be used to come up with further and different innovations. That is, how much do past, even unsuccessful, research efforts contribute to future successes? Anecdotal evidence is mixed, and serious empirical work is completely missing. Similarly, the function $g(\lambda)$ also needs to be measured. Are we all drawing from the same Platonic urn of ideas? Does schooling allow us, at least, to sample without replacement, so that more people means faster drawing of new ideas? Or are we, instead, sampling from the same urn with replacement, in which case more and more useless duplicates are drawn as λ increases? Or, maybe, adding more people increases the number of original ideas in the urn from which we sample, making $g(\lambda)$ increase quite fast? These are important empirical and theoretical questions that current research has not yet considered.

Turning to the policy implications of our results: if $\phi\lambda$ is held fixed, the quality of ideas produced remains unchanged. This is not the social optimum: generally we will want to take advantage of the increased λ to allow some marginal ideas to enter the market. A simple rule of thumb is to observe that for large λ a linear $M(\rho)$ implies that protection should be reduced by about half the increase in scale of market. Thus, the simple rule of thumb would be that if the size of market doubles, the amount of protection should be reduced by about 1/3—that is, a 50% increase in 2/3 the level of protection is half of the 100% increase in the size of the market. Taking a real interest rate of 2% per year, observe that $1-\phi = e^{-rT}$, where T is the length of IP protection. Using $\log(1+x) \approx x$, this gives $\Delta T \approx 50\Delta\phi$.

For example, the G7 nations account for about 2/3rds of world GDP. If we think of the intellectual property changes in the WTO as extending the protection that exists in the G7 nations to the rest of the world, this suggests a reduction in protection by about 1/6. This means approximately an 8-year reduction in term. Similarly, if the world economy is growing at 2% a year, a simple rule of thumb would be to reduce protection terms by 1% per year, or about 6 months per year.

A paradigmatic case is that of popular music. Forty years ago, at the time of Elvis Presley and the Beatles, new recordings selling a million units were considered exceptional successes and awarded "gold records," while in the current times a successful record sells easily 10 to 20 million copies. The effective size of the market has, therefore, increased by at least a factor of ten. At the same time, advances in recording and digital technologies have

reduced the fixed cost required to produce a new record to about one-fifth of its earlier level. This suggests that the socially optimal length of copyright protection should have dropped by about a factor of 25. Unfortunately, in the case of copyright, terms have been moving in the opposite direction; copyright terms have grown by a factor of about four since early in the twentieth century. This means that, at least for recorded music, they currently are on the order of 100 times longer than they should be. A similar calculation can be performed for books and movies. Consider the fact that, since the beginning of the past century, world GDP has grown by nearly two orders of magnitude. It is reasonable to argue that the size of the market for books and movies must have grown by at least as much, as literacy has surged and the availability of playing devices has increased more than proportionally due to the dramatic drop in their relative prices. Hence, if the copyright term of 28 years at the beginning of the twentieth century was socially optimal, the current term should be about 6 months, rather than the current term of approximately 100 years. This gives a ratio of 200 between the actual copyright terms and their socially optimal value.

Our results are relevant for the debate on the impact of IP harmonization policies on developing countries. Romer [1994], among others, has pointed out that in the presence of fixed costs the welfare loss from tariff protection may be orders of magnitude larger than the usual Harberger's triangle. The same logic applies to the welfare losses due to IP protection. The point to notice here is that monopoly prices due to patents and copyright have the same effect as taxes on *ex post* profits and lead to the noncreation of a new good or the nonadoption of a new production process when the indivisibility is relevant.

References

[1] Acemoglu, A. and F. Zilibotti (1996), "Was Prometheus Unbound by Chance? Risk, Diversification and Growth," *Journal of Political Economy,* **105**, 709–751.

[2] Akerloff, G., K. Arrow, T. Bresnahan, J. Buchanan, R. Coase, L. Cohen, M. Friedman, J. Green, R. Hahn, T. Hazlett, C. Hemphill, R. Litan, R. Noll, R. Schmalensee, S. Shavell, H. Varian, and R. Zeckhauser (2002), *Amici Curiae in Support of Petitioners in the Supreme Curt of the United States, Eldred versus Ashcroft.*

[3] Boldrin, M. and D. K. Levine (1999), "Perfectly Competitive Innovation," mimeo, University of Minnesota and UCLA, November.

[4] Boldrin, M. and D. K. Levine (2002), "The Case Against Intellectual Property," *American Economic Review (Papers and Proceedings)* **92**, 209–212.

[5] Boldrin, M. and D. K. Levine (2004a), "IER Lawrence Klein Lecture: The Case Against Intellectual Monopoly," *International Economic Review,* **45,** 327–350.

[6] Boldrin, M. and D. K. Levine (2004b), "The Economics of Ideas and Intellectual Property," *Proceedings of the National Academy of Sciences,* forthcoming.

[7] Diwan, I. and D. Rodrik (1991), "Patents, Appropriate Technology, and North- South Trade," *Journal of International Economics,* **30,** 27–48.

[8] Eckstein, Z. and E. Nagypal (2004), "U.S. Earnings and Employment Dynamics 1961–2002: Facts and Interpretations," mimeo, University of Minnesota and Northwestern University, January.

[9] Gallini, N. (1992), "Patent Policy and Costly Imitation," *RAND Journal*, **23,** 52–63.

[10] Gilbert, R. and C. Shapiro (1990), "Optimal Patent Length and Breadth," *RAND Journal*, **21,** 106–112.
[11] Grossman, G. M. and E. Helpman (1991), "Trade, Knowledge Spillovers and Growth," *European Economic Review (Papers and Proceedings)*, **35,** 517–526.
[12] Grossman, G. M. and E. Helpman (1994), "Endogenous Innovation in the Theory of Growth," *Journal of Economic Perspectives*, **8,** 23–44.
[13] Grossman, G. M. and E. Helpman (1995), "Technology and Trade," in G. Grossman and K. Rogoff, eds., *Handbook of International Economics*, vol. III, North Holland.
[14] Grossman, G. M. and E. L. Lai (2002), "International Protection of Intellectual Property," NBER Working Paper 8704.
[15] Grossman, G. M. and E. L. Lai (2004), "International Protection of Intellectual Property," mimeo, Princeton.
[16] Hall, B. (2001), "NBER Patents Data File," http://elsa.berkeley.edu/users/bhhall/pat/datadesc.html.
[17] Hall, B., A. Jaffe, and M. Tratjenberg (2004), "Market Value and Patents Citations," *RAND Journal of Economics*, forthcoming.
[18] Hart, O. D. (1979), "On Shareholder Unanimity in Large Stock Market Economies," *Econometrica,* **47,** 1057–83.
[19] Harhoff, D., F. M. Scherer, and K. Vopel (1997), "Exploring the Tail of Patented Invention Value Distributions," discussion paper FS IV 97-27, WZB, Berlin.
[20] Jones, C. (2004), "Growth and Ideas," Prepared for *Handbook of Economic Growth.*
[21] Landes, W. M. and R. A. Posner (2003), *The Economic Structure of Intellectual Property Law*, Harvard University Press.
[22] Lanjouw, J. (1993), "Patent Protection: Of What Value and How Long?" NBER Working Paper 4475.
[23] Lanjouw, J., A. Pakes, and J. Putnam (1998), "How to Count Patents and Value Intellectual Property: The Uses of Patent Renewal and Application Data," *Journal of Industrial Economics,* **46,** 405–432.
[24] Leibowitz, S. and S. Margolis (2003), "Seventeen Famous Economists Weigh in on Copyright: The Role of Theory, Empirics, and Network Effects," mimeo, University of Texas at Dallas.
[25] Lo, S. (2003), "Strengthening Intellectual Property Rights: Experience from the 1986 Taiwanese Patent Reforms," mimeo, UCLA.
[26] Makowski, L. (1980), "Perfect Competition, the Profit Criterion and the Organization of Economic Activity," *Journal of Economic Theory,* **22**, 222–42.
[27] Maurer, S. M., and S. Scotchmer (2002), "The Independent Invention Defense in Intellectual Property, *Economica,* forthcoming.
[28] Pakes, A. (1986), "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, **54**, 755–784.
[29] Park, W. and D. Lippholdt (2003), "The Impact of Trade-Related Intellectual Property Rights on Trade and Foreign Direct Investment in Developing Countries," OECS, http://www.american.edu/cas/econ/faculty/park/TD-TC-WP-2003-42-final.pdf.
[30] Romer, P. M. (1990), "Endogenous Technological Change," *Journal of Political Economy,* **98**, S71–S102.
[31] Romer, P. (1994), "New Goods, Old Theories, and the Welfare Costs of Trade Restrictions," *Journal of Development Economics*, **43**, 5–38.
[32] Sampat, B. N. and A. A. Ziedonis (2002), "Cite Seeing: Patent Citations and the Economic Value of Patents," mimeo, Georgia Institute of Technology and University of Michigan, November.
[33] Scotchmer, S. (1991), "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law," *Journal of Economic Perspectives,* **5,** 29–41.

[34] Silverberg, G. and B. Verspagen (2004), "The Size Distribution of Innovations Revisited: An Application of Extreme Value Statistics to Citation and Value Measures of Patent Significance," MERIT working paper 2004-021, Maastricht University.
[35] Taylor, M. S. (1993), "TRIPS, Trade, and Technology Transfer," *Canadian Journal of Economics,* **26,** 625–637.
[36] Taylor, M. S. (1994), "TRIPS, Trade, and Growth," *International Economic Review,* **35,** 361–381.
[37] U.S. Copyright Office (2001a), *A Brief History and Overview*, Circular 1a, http://www.copyright.gov/circs/circ1a.html.
[38] U.S. Copyright Office (2001b), *Copyright Registration for Sound Recordings*, Circular 56, http://www.copyright.gov/circs/circ56.html.
[39] Yang, G. and K. E. Maskus (2001), "Intellectual Property Rights, Licensing, and Innovation in an Endogenous Product Cycle Model," *Journal of International Economics,* **53,** 169–187

Appendix: Data

Book Revenue. We collected all the titles, ISBNs, and sale prices listed at www.amazon.com for the query *hardcover fiction books* and for the two publication periods of September 2003 and September 2004. The sales data are from the Ingram stock statistics, automatic telephone line at 615-213-6803. The Ingram stock statistics system gives the following statistics for each ISBN punched in: "total sales this year," "total sales last year," "total current unadjusted demand," "total last week demand." Total revenue for each book is calculated using the total sales data from Ingram and the November 2004 sales price listed at www.amazon.com. Ingram is a large book distributor, and is generally thought to generate roughly 1/6 of all book sales. It should be noted that the sales prices at www.amazon.com are changing over time, most often decreasing, so we might have underestimated the revenue during the first year for books published during September 2003. Because of the large number of observations, we do not reproduce the data here, but they are available at http://www.dklevine.com/data.htm.

Copyright Time Series. The basic source of the copyright registration time series is the annual report of the Copyright Office from 2000, which can be found at http://www.copyright.gov/reports/annual/2000/appendices.pdf. This also includes the breakdown of registrations by type for 2000. Population data for 1901–99 are from the U.S. Census Bureau at http://www.census.gov/population/estimates/nation/popclockest.txt. Data prior to 1901 are from http://www.census.gov/population/censusdata/table-2.pdf. The two sources have a slight discrepancy for 1900 population with the former source reporting 76,094,000 and the latter (which we used) 76,212,168. The year 2000 data are from the 2000 census. Literacy rates are from http://www.arthurhu.com/index/literacy.htm. The data we used can be found at http://www.dklevine.com/data.htm.

Patent Time Series. R&D Expenditures by Sectors: National Patterns of R&D Resources: 2002 Data Update, Table D, National Science Foundation. GDP.: National Income and Production Account, Table 1.1.5, Bureau of Economic Analysis. Population: 1953–1959: Population Estimates Program, Population Division, U.S. Census Bureau, release date: April 2000. 1960–2002: U.S. Census Bureau, Statistical Abstract of United States, 2004–2005.

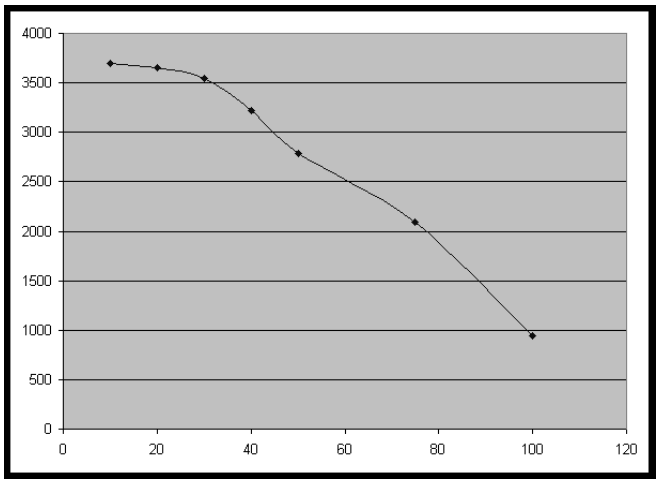

Figure 1: U.S. Income Distribution 2001

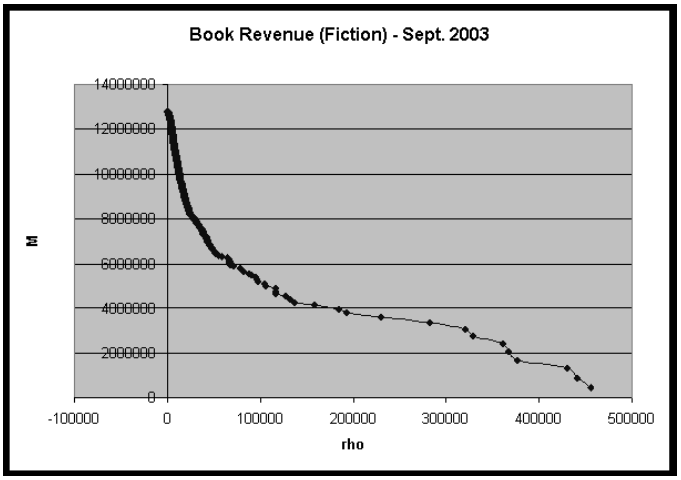

Figure 2

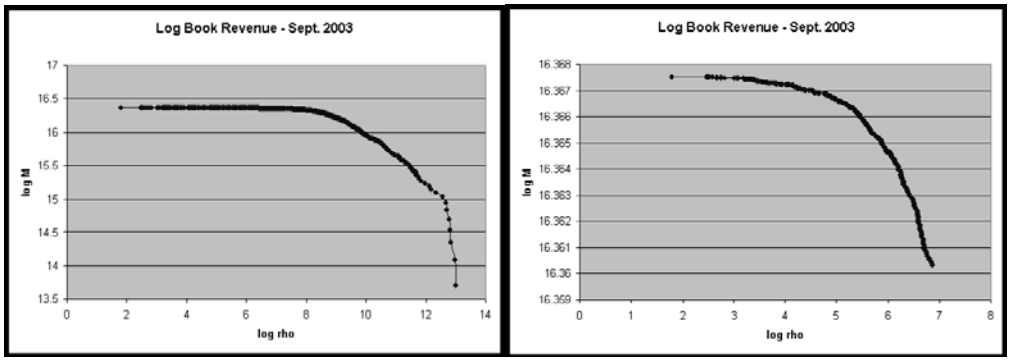

Figure 3

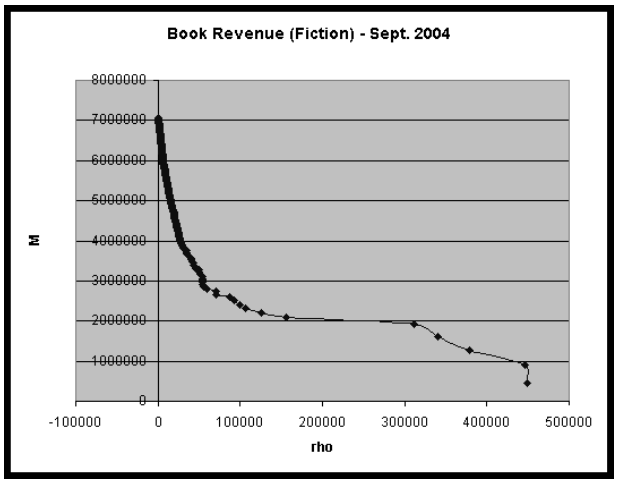

Figure 4

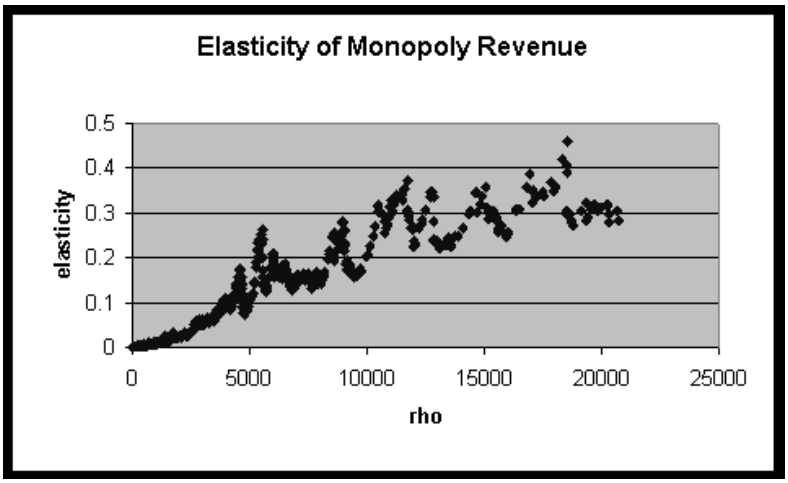

Figure 5

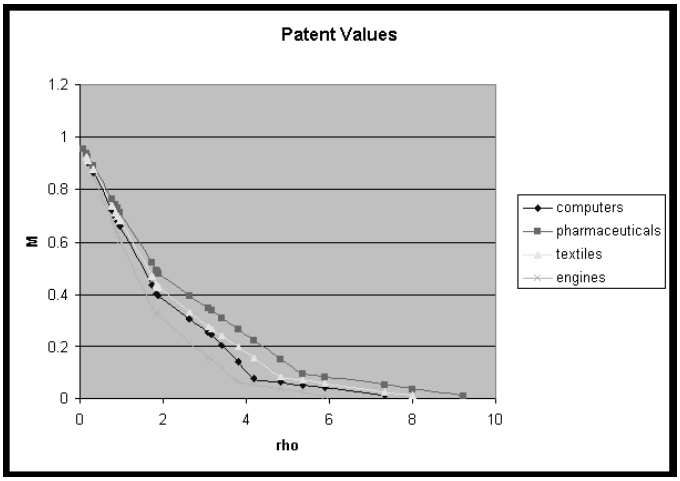

Figure 6

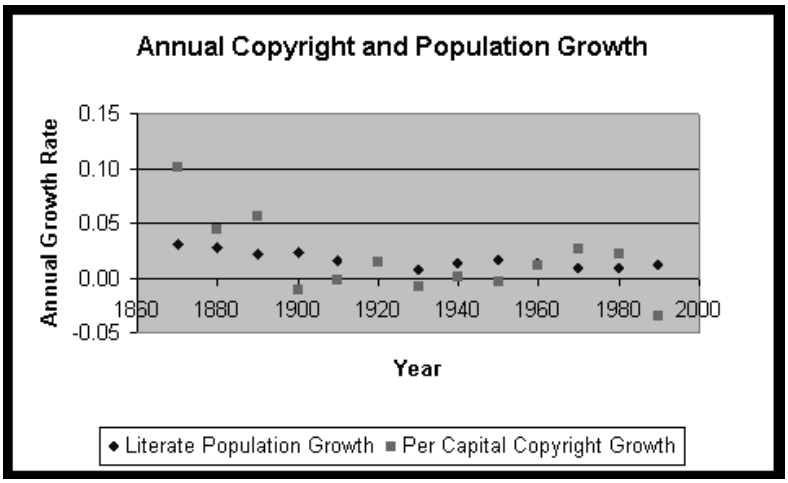

Figure 7

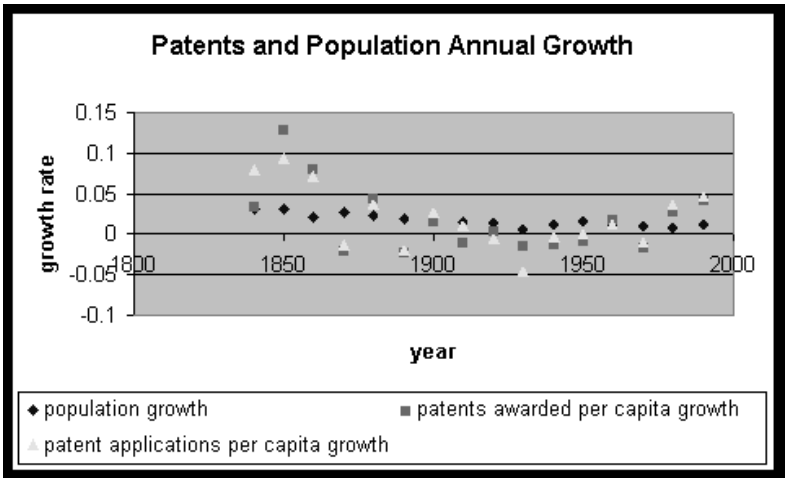

Figure 8