IP AND MARKET SIZE

MICHELE BOLDRIN AND DAVID K. LEVINE

Department of Economics, University of Minnesota, the Federal Reserve Bank of Minneapolis and CEPR; UCLA and the Federal Reserve Bank of Minneapolis

ABSTRACT. Intellectual property (IP) protection involves a trade-off between the undesirability of monopoly and the desirable encouragement of creation and innovation. Optimal policy depends on the quantitative strength of these two forces. We give a quantitative assessment of IP policy. We focus particularly on the scale of the market, showing that as it increases, due either to growth, or to the expansion of trade through treaties, IP protection should be reduced.

Keywords: Property, International Trade, Innovation, Harmonization, Monopoly.

JEL Classification: X15; X16

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 Dallas Federal Reserve Bank, the faculty seminar at the Chinese University of Hong Kong, the Theory Workshop at ASU, the CSIP seminar at the San Francisco Federal Reserve Bank, and from discussions with Kyle Bagwell, Gene Grossman, Chad Jones, Pete Klenow, Paul Romer, Suzanne Scotchmer and Daniel Wilson. Thanks are also due to Hengjie Ai research assistance in collecting and analyzing data.

Corresponding Author: David K. Levine, Department of Economics, UCLA, Los Angeles, CA 90095, USA. Phone/Fax: 310-825-3810. Email: david@dklevine.com.

First Version: 2nd December, 2003, This Version: 22nd June 2005.

1. INTRODUCTION

In this paper we look at the general equilibrium interactions that determine the optimal level of protection for Intellectual Property (IP). There is a large literature that explores the qualitative aspects of optimal IP policy, and a significant empirical literature that attempts to measure such things as the value of patents. There is, however, little connection between the two. Our goal is to use a relatively standard model of IP, based on that of Grossman and Lai [2004], and examine the quantitative implications of existing empirical results.

Profitability of an innovation depends upon three factors: the initial cost of discovery, the elasticity of demand, and the size of the market; all these elements vary widely and unsystematically across innovations. We focus primarily on market size as it is substantially easier to measure than the other two and, since current patent and copyright legislation was first introduced, it has grown steadily and substantially. Finally, the WTO-TRIPS agreement has put the relationship between market size and harmonization of IP at center stage.

Optimal policy involves a trade-off between increasing the monopolistic distortion on inframarginal ideas, and increasing the number of marginal ideas. As the scale of the market increases, depending on elasticity, it may 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. In this case the optimal policy should reduce the length of protection as the scale of the market increases. Our analysis of a variety of data show that this is empirically relevant case.

We utilize a standard model in which IP protection is socially beneficial.[1] Ideas are created subject to a 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 in a market of unit size. We focus mainly on the case in which the private return is proportional to the social return to 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.

[1]We have examined the shortcomings of the "standard model" in Boldrin and Levine [1999, 2002, 2004, 2005] where we argue that IP is not generally socially beneficial.

Our 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 [2004]. Both they and we show that as the total monopoly revenue function has increasing (decreasing) elasticity, then optimal protection locally decreases (increases). From a theoretical perspective, 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, and examining balanced growth paths. Since they have already provided this interpretation, we simply note that this procedure is equally valid for our model. 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. We use a disaggregated model in which ideas are heterogeneous both in their quality and in their cost of creation.[2] The latter gives us a useful tool for quantitative analysis, as the distribution of private returns from ideas can be estimated from available data.

We also examine the (skilled) labor demand implications of the model. It turns out that when elasticity is constant or decreasing, the demand for labor increases by more than the increase in market scale. This ultimately leads to a binding labor constraint. Once the labor constraint binds, the only effect of increasing market scale is to increase rents accruing to scarce R&D labor. Reducing IP protection reduces these rents – with no consequence for welfare – and lowers the monopoly distortion on the innovations that are all going to be created anyway. Hence we show, regardless of elasticities, that sufficiently large market size implies decreasing protection with increasing market size. Our labor supply analysis also enables us to examine data on labor demand in the R&D sector as a second source of information on the elasticity of total monopoly revenue.

We conclude that both theoretical reasoning and empirical evidence strongly suggest that we are in a region where the optimal level of protection decreases with the scale of the market. In addition to considering increases in the scale of the market through growth, we reexamine the Grossman-Lai exercise of "harmonization" and North-South trade from a quantitative perspective. Our main finding is that the North should reduce protection as a result of harmonization.[3] In the case of two (or more) countries of equal size, because some of the benefits of higher IP protection are received by

[2] Their production function implies a particular distribution of private returns. We show how to make this connection, and that the reduced form of the two models are equivalent.

[3] This contradicts one of Grossman-Lai's stated theoretical results; we believe that their finding is due to an algebraic error.

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 everywhere. However, this argument applies only to countries of equal size. When the countries, two or more, are of unequal size, smaller countries tend 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 countries. 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 limits to lie in between the larger protection of the larger country and the smaller protection of the smaller countries.

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 of $\Re^n$.[4] 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 population of size λ of agents, which measures 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 non-decreasing; without loss of generality let $g(1) = 1$. 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, are possibilities considered here.

Once an idea is created, it may be reproduced at no cost and without limit. If the input of labor $y(\omega)$ is below the threshold, that is $y(\omega) < h(\omega)$, ideas of type ω cannot be reproduced. If $y(\omega) \geq h(\omega)$, then aggregate reproduction and consumption is $x(\omega) \geq 0$, and consumption per capita is $z(\omega) = x(\omega)/\lambda$.

[4] Actually, any topological measure space will do. The importance of treating ideas as diverse rather than using a production function for knowledge has been emphasized by Scotchmer [1999] for example.

Returns from ideas are uncertain at the time the invention decision is made; therefore it is the *ex ante* expected return on an idea and not its *ex post* realization that matters for the decision to invent. For concreteness, imagine that z units of an idea with characteristics ω have utility net of production costs to a representative consumer of $u(z,\omega)$ with probability $p(\omega)$, while with probability $1-p(\omega)$ the idea has no utility at all and is not produced. Normalize $u(0,\omega)=0$, and assume that $p(\omega), u(z,\omega)$ are continuous in ω with the latter also continuous and non-decreasing in z and, at least up to a limit $Z(\omega)$, smooth and strictly increasing.[5]

Let $v(z,\omega)=p(\omega)u(z,\omega)$ be expected utility; let $\bar{z}(\omega)$, possibly infinite, be the least value such that $v(\bar{z}(\omega),\omega)=0$; and assume

$$\lim_{z\to\bar{z}(\omega)} v(z,\omega)=v^C(\omega)<\infty.$$

Since $v(z,\omega)$ is bounded, $zv_z(z,\omega)\to 0$ as $z\to\bar{z}(\omega)$, that is, per capita revenue[6] falls to zero as per capita consumption grows to the maximum. 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 spent on activities that take place outside of the idea sector, the marginal utility of which we normalize to one. Denoting with L the individual endowment of time, the aggregate feasibility constraint in an economy of size λ is

$$\lambda L \geq \int y(\omega)g(\lambda)\eta(d\omega)$$

and individual utility is

$$\int v(z(\omega),\omega)g(\lambda)\eta(d\omega)+L-\frac{1}{\lambda}\int y(\omega)g(\lambda)\eta(d\omega).$$

Profit maximization and efficiency require $y(\omega)=h(\omega)$ for all ideas for which $x(\omega)>0$ and $y(\omega)=0$ otherwise. Obviously, no good would be produced in this economy absent patent protection.

[5]Returns can also be uncertain because the cost of an idea is uncertain. It is easy to check that the representation of uncertainty suggested here can be applied to cost as well as to benefit.

[6]Profits, strictly speaking, include the fixed cost of producing the idea, while revenues should exclude the variable cost of producing z from the idea. The key variable in our analysis is the intermediate notion of revenue net of the variable cost of production; for convenience we refer to this simply as revenue.

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.[7] While the patent lasts, the innovator 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 freely make copies, and output and consumption jump to $\bar{z}$ while price falls to marginal cost and revenue falls to zero.[8] An idea is produced if, given the patent length ϕ, the prospective monopolist finds it profitable to overcome the indivisibility.

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 sells $z(\omega)$ units of output to each of the λ consumers receives revenue[9] $\lambda z(\omega) v_z(z(\omega), \omega)$, which is assumed to have a unique maximum at $z^M(\omega)$, and pays the cost $wh(\omega)$. The ratio of (per capita) private value to the total cost of innovation is $\rho(\omega) = z^M(\omega) v_z(z^M(\omega), \omega)/h(\omega)$. In fact $\rho(\omega)/w$ represents one plus the rate of return on investment that 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

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

No ideas with a $\rho(\omega)$ lower than $\underline{\rho}$, and all ideas with a $\rho(\omega)$ above $\underline{\rho}$ will be produced in the patent equilibrium. 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 monotone relation between the private return $\rho(\omega)$ of an idea and its social return; hence ideas of high social return

[7] We do not model the "patent race" by which patent is awarded; just assume that, for each of the $\eta(\omega)$ ideas with characteristics ω, a particular individual is awarded a "patent."

[8] We assume there are no competitive rents after the patent expires; as pointed out in Boldrin and Levine [1999], inventors generally do earn positive competitive rents.

[9] If marginal cost of producing z is increasing, then we have assumed that the producer surplus does not go to the monopolist, but to other factors of production. This assumption seems the most empirically relevant and is for concreteness, playing no important role in the analysis.

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 cutoff ρ we may 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, in an economy of unit size. 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 *strong return neutrality*. This occurs when the ratios of social values to private return $\nu^M(\omega)$ and $\nu^C(\omega)$ are both constant. When all ideas are identical from the point of view of consumers, which is a common assumption in the literature, return neutrality also holds.

From a formal point of view, 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 Σ 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 $\overline{f}(\rho)$, 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 $B \subset \Sigma$. By *return neutrality* we mean that $\overline{\nu}^M(\rho), \overline{\nu}^C(\rho)$ are constant.

For the remainder of the paper, we will assume return neutrality. It is worth noting briefly what happens when this assumption fails. There are two possibilities. If goods with lower private return have also lower social value, in the sense that $D\overline{\nu}^M(\rho) > 0$ and/or $D\overline{\nu}^C(\rho) > 0$, common sense and simple calculations show that this further strengthens the argument that the length of protection should decline with the scale of the market. The opposite case $D\overline{\nu}^M(\rho) < 0, D\overline{\nu}^C(\rho) < 0$ might seem to have the opposite effect, reinforcing the case for increasing IP protection. In this case, however, 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. Of course it can be argued that strong IP protection is needed because that is the only way to get marginal ideas – that is, ideas of high social merit – produced. But this argument assumes that the only policy instrument is IP protection. If it were really the case in practice that privately valuable innovations have little or no social value, and vice versa, then almost any form of government intervention other than IP would be sensible. Publicly sponsored research projects, auctioning of production rights, or subsidies for innovators producing the socially valuable ideas would all make sense; IP would not.

3. Optimal IP Protection under Return Neutrality

We first ask how socially optimal protection $\hat{\phi}(\lambda)$ depends on market size.

Proposition 3.1. *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* $\lambda > \overline{\lambda}$. *Further, when* $\hat{\phi}(\lambda) < 1$, *in a neighborhood of* $\rho = 1/\lambda\hat{\phi}(\lambda)$, *the following three cases hold. (I)* $\Upsilon'(\rho) > 0$ *implies* $\hat{\phi}(\lambda)$ *is unique and strictly* *decreasing; (II)* $\Upsilon'(\rho) = 0$ *implies* $\hat{\phi}(\lambda)$ *is unique and constant; and (III)* $\Upsilon'(\rho) < 0$ *and* $\hat{\phi}(\lambda)$ *unique*[10] *implies* $\hat{\phi}(\lambda)$ *is strictly increasing.*

The details of the proof are in Appendix 1. Here we provide a sketch. Basically, in the proof we treat the global (first part of the proposition) and local (second part) cases separately. The local case is analyzed by dividing the first order condition for a welfare maximum by $M(1/\phi\lambda)$ to get a condition we refer to as the NOC:

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

The NOC has the same qualitative properties as the first order condition: it has the same zeros, the same sign on the boundary and $NOC_\phi(\lambda,\phi) < 0$ is sufficient for a zero to be a local maximum. In Appendix 1 we differentiate the NOC and verify the second order and boundary conditions and apply the implicit function theorem to prove the local results. The global case requires examination of labor demand, which we further discuss below.

The Production Function Approach. Grossman and Lai [2004] adopt a production function approach in which $Q = F(H,L)$ homogeneous ideas are produced by a fixed amount of human capital, H, and labor, L, with F 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 return production function. Observe that when $w = 1$ the total labor cost of producing Q ideas is $f^{-1}(Q)$, and the cost of the marginal idea 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 to the marginal idea produced by the Lth unit of labor input is revenue divided by the cost of producing it, 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 may easily derive that the elasticity Υ of M is the same as Grossman and Lai's elasticity of research output with respect to labor.[11]

Grossman and Lai focus on the CES family of production functions. In the case of the Cobb-Douglas production function elasticity Υ is constant. Less substitutability between human capital and labor implies increasing elasticity, and conversely.

Our distributional theory gives some insight into what this elasticity is likely to look like. Under the plausible assumption that there are ideas so

[10] In this case we cannot guarantee that the second order condition is satisfied, so we must rule out the possibility that $\hat{\phi}(\lambda)$ has multiple values.

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

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 when ρ is small. 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 high degrees of substitutability in which elasticity is globally decreasing or even constant.

In this same direction, most common distributions give rise to increasing elasticity. This is true for the exponential, normal, lognormal, and truncated Pareto distributions. By way of contrast, if $\mu(\rho)$ is globally Pareto, then elasticity is constant. In this case $M(\rho)$ 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, we would not expect it to hold globally and, certainly, not for ρ close to zero. As we shall see, the data suggest that modeling $M(\rho)$ as a linear function is a plausible approximation; modeling it as a Pareto is not.

Labor Demand. Our argument that for sufficiently large scale of market optimal protection must be decreasing is based on our analysis of labor demand. Setting $\ell(\underline{\rho}) = \int_{\underline{\rho}}^{\infty}\mu(\rho)d\rho$, labor demand is given by

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

from which, letting $\mathfrak{E}$ denote the elasticity operator, we have

$$\mathfrak{E}[L^D(\lambda)] = \mathfrak{E}[g(\lambda)] - \mathfrak{E}[\ell(\rho)].$$

Depending on which assumptions one makes about $g(\lambda)$, the first factor ranges from zero to any large positive number. For example, if one takes the production function approach, then $g(\lambda)$ can be identified with aggregate human capital H. To the extent this is constant, $\mathfrak{E}[g(\lambda)] = 0$; if, instead, $H = h\lambda$, then $\mathfrak{E}[g(\lambda)] = 1$. In models of growth and innovation due to externalities or increasing returns, such as Grossman and Helpman [1991, 1994, 1995] or Romer [1990], $g(\lambda)$ is assumed to increase faster than λ, hence $\mathfrak{E}[g(\lambda)] > 1$. A benchmark case is that in which each individual draws her own ideas from the same urn, either 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 $\mathfrak{E}[g(\lambda)] = 1$; if sampling is with replacement then $\mathfrak{E}[g(\lambda)] \leq 1$.

As for the second factor, notice first that the demand for labor is linked to the total revenue function by the following relation

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

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

$$\mathfrak{E}[L^D(\lambda)] = \mathfrak{E}[g(\lambda)] + \zeta + 1.$$

Notice that when $\mathfrak{E}[g(\lambda)] > 1-\zeta$, the elasticity of labor demand is predicted to be larger than two, hence the elasticity of per capita labor demand is greater than one. More generally, since $\mathfrak{E}[g(\lambda)] \geq 0$, we have $\mathfrak{E}[L^D(\lambda)/\lambda] > 0$. In other words, in the data, as the size of the economy grows, the *share* of workers in the idea sector grows as well. This is the intuition underlying the global statement in our first proposition: if elasticity is not increasing, then eventually the labor constraint must bind. The next proposition, proven in Appendix 1, shows how to extend from the case of constant elasticity to decreasing elasticity of the total monopoly revenue.

Proposition 3.2. *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, $\Upsilon_1(\rho) = \Upsilon_2(\rho)$) 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 $\mathfrak{E}[\ell_1(\rho)] > \mathfrak{E}[\ell_2(\rho)]$.*

(3) *As the elasticity of total revenue goes from increasing, to constant, to decreasing, the elasticity of the associated labor demand functions increases monotonically.*

In plain words: a revenue function with decreasing elasticity implies an elasticity of labor supply even larger than that of a constant elasticity revenue function, which we have shown to be at least one in practice. Playing this backward: should the empirical elasticity of per capita labor supply with respect to market size be smaller than one, then the associated total revenue function must display increasing elasticity. Per capita labor in the idea sector growing faster than the scale of market is consistent with increasing elasticity of total monopoly revenue, because $\mathfrak{E}[g(\lambda)]$ can be large, which is independent of the elasticity of monopoly revenue. However, if per capita labor grows more slowly than the size of the market, we must rule out both constant and decreasing elasticity.

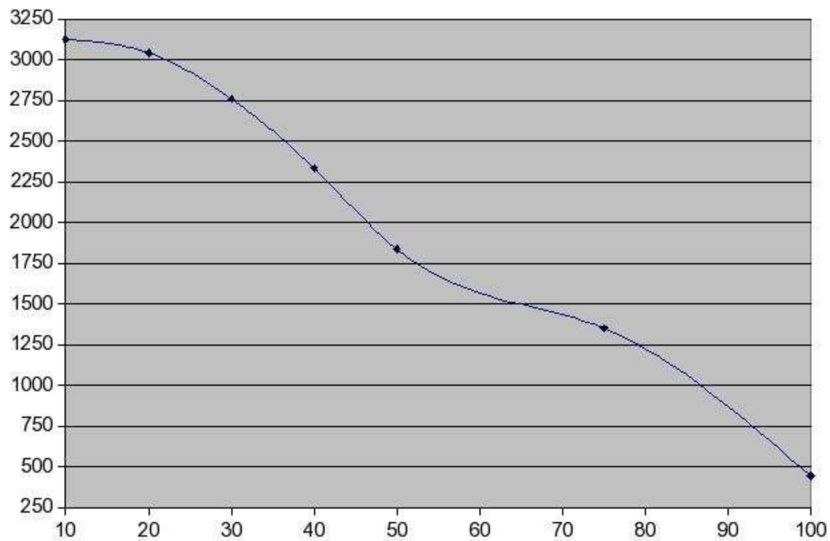
4. Empirical Analysis of Total Monopoly Revenue

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 his own ideas and his own opportunity costs of engaging in innovative activity. 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" $\underline{\rho}$.

An issue arises at this point. In the available data we observe revenues, $zu_z(\omega,z)$, not returns, ρ. Further, it is hard to observe directly either the opportunity cost, w, of each inventor or the labor cost $h(\omega)$ of ideas. Hence we need to assume that, for all the ideas ω in the data set, the product $wh(\omega)$ is a constant. This ensures that returns are proportional to revenues. If our data sets contained ideas produced in very different sectors, this assumption would be absurd. To avoid this we will try to restrict attention to sets of goods that are relatively homogeneous, so that it is reasonable to assume that $wh(\omega)$ is roughly constant within each set. A second issue also arises when going to the data. We set $v(\omega,z) = p(\omega)u(\omega,z)$, where $1 - p(\omega)$ is the probability that the idea turns out to have no value *ex post*. We do not observe $v_z(\omega,z) = p(\omega)u_z(\omega,z)$, but rather just $u_z(\omega,z)$, which is the market price of a copy of idea ω. As long as $p(\omega)$ does not depend on ω, this means that we will overestimate revenues but will correctly compute elasticities. So we must assume that $p(\omega)$ is constant. We have also assumed that *ex ante* there is only one "successful" outcome; if there is a large variation in outcomes, this too will pose a problem for data analysis.

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. 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. Let ρ denote personal income here; the corresponding $M(\rho)$ from the Current Population Survey data on 2001 income is plotted in Figure 4.1.

FIGURE 4.1. U.S. Income Distribution



Horizontal Axis: 2001 Individual Income in \$1000 USD (ρ)
Vertical Axis: Cumulative Income ($M(\rho)$)
Source: Current Population Survey

Eyeballing is enough to realize 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.

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 available. Although we do not have data on lifetime income of individual authors, we do have data on the revenue generated by individual book sales. We ignore the fact that it is costly to produce books once they are written, which is irrelevant to our ends insofar as the cost of producing each copy of a book is independent of the number of copies produced and sold. In summary, our assumptions are

- The opportunity cost $wh(\omega)$ of writing books is constant.
- Expected revenues from the sale of a "successful" book are perfectly anticipated, and the probability of failure does not depend upon ω.
- The marginal cost of producing books is small relative to sales price.

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 fiction books published in March and September of 2003 and 2004, respectively; our samples range between 1,200 and 1,300 books for each of these four months. The details of the data collection procedure can be found in Appendix 2. Figure 4.2 shows $M(\rho)$ computed on the basis of the September 2003 data.

FIGURE 4.2. _

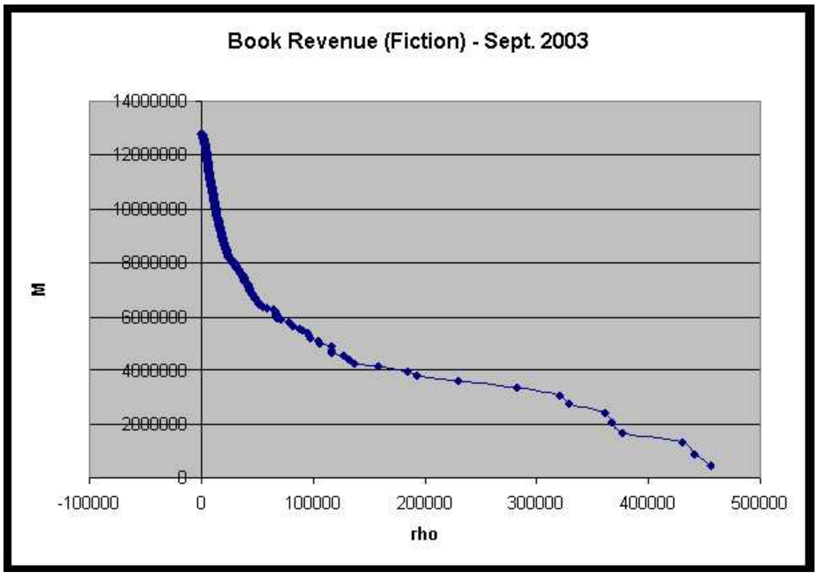
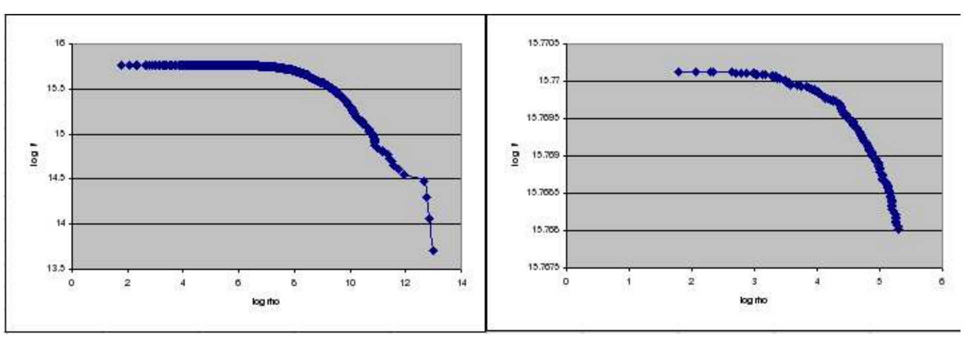


Figure 4.3 shows a plot on logarithmic axes, including a close-up to illustrate more clearly the increasing nature of the elasticity on both ordinary and logarithmic axes.

FIGURE 4.3. Logarithmic Book Revenue (Fiction) – Sept. 2003



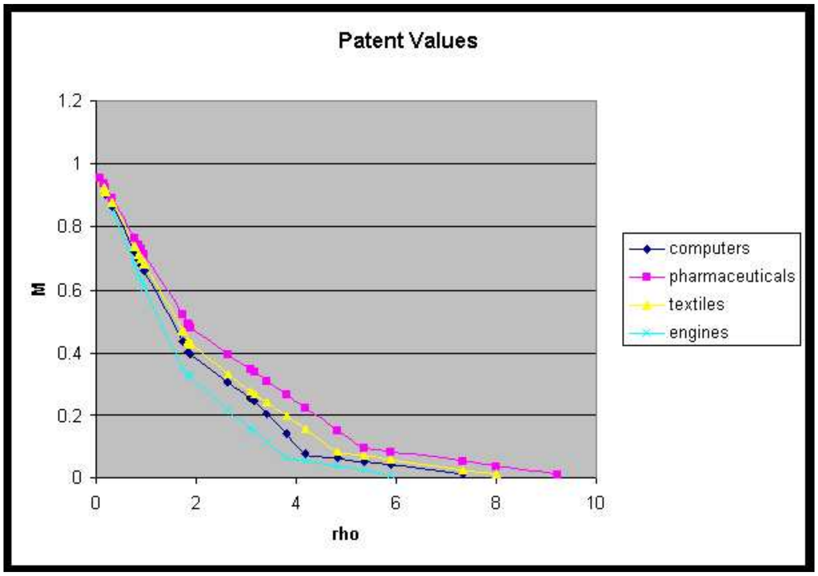
The data for the other months, not reported but available, yield extremely similar results.

Three comments are in order. First, 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 second striking feature is the discontinuity between roughly \$150,000 and \$300,000 in revenue.[12] 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 two-thirds of all book revenues. This is considerably more concentrated than we find in our data – but certainly reflects a strong discontinuity. These books appear to be predominately by "big name" authors, 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 third fact is how small the indivisibility $wh(\rho)$ may be for writing and publishing fiction; in September 2003, 1,181 books, out of a total of 1,223, 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 the other months are similar. In the same 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.

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*. If we disaggregate by industry, it is at least plausible that the fixed cost of the innovation is not systematically related to the realized revenues. We use 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 4.4.

[12]The sales data are from a single distributor, Ingram, constituting about one-sixth of the book market, so total revenues would be about six times this number.

FIGURE 4.4. _



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. Numerical estimates can be found in Figure 4.5. This reports for each industry and for increasing values of ρ, the elasticities evaluated at the midpoint of each segment of the linear spline. The number in square brackets is the corresponding value of $-\rho M_i'(\rho)$.

FIGURE 4.5. Elasticities

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]

With the exception of the highest category of ρ for computers, elasticities are increasing everywhere. The values of $-\rho M_i'(\rho)$ are also relevant because the same ϕ applies across sectors. Hence the aggregate distribution is $M(\rho) = \sum_i M_i(\rho)$, where i indexes industries. Unfortunately, the fact that each $M_i(\rho)$ function has increasing elasticity does not imply that this is true for $M(\rho)$. However, if $M_i'(\rho)$ is increasing, then the corresponding elasticity is increasing as well, and increasing $-\rho M_i'(\rho)$ is a condition that does aggregate. While not always increasing, $-\rho M_i'(\rho)$ is increasing in the relevant range, that is, at lower values of ρ, for all i. This implies the elasticity of $M(\rho)$ is also increasing, at least for values of ρ near the threshold, which is what matters.

Our findings for patents appear to accord well with the existing empirical literature. 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 ability of various empirical distributions, including the Pareto, to fit the data and find that a two-parameter lognormal distribution provides the best fit. Silverberg and Verspagen (2004) use a variety of different data sources from both Europe and the U.S.A. and two different measures of ρ (citations and monetary values). 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 econometric literature on the value of patents, stemming from the paper of Pakes [1986] (see Hall, Jaffe, and Tratjenberg [2004] for a recent update and new results), seems to almost unanimously find that the appropriate distribution is a log-normal or an exponential, for both of which the elasticity of the total revenue function is increasing.

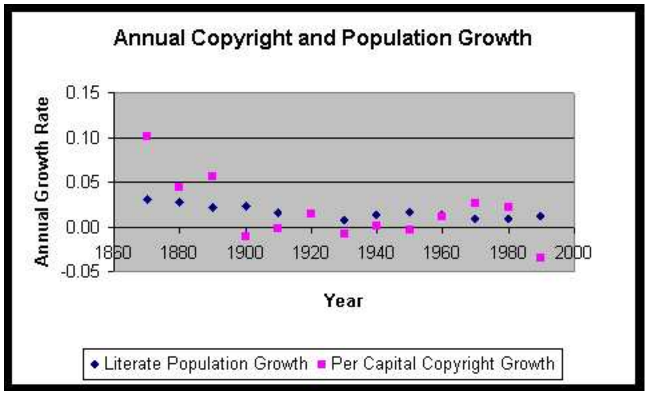
5. Empirical Analysis of Labor Demand

As we have seen, there is a close connection between the elasticity of total monopoly revenue and labor demanded by the ideas sector. Here we exploit that relationship to get a second source of information about whether the elasticity of total monopoly revenue is increasing or decreasing.

Copyright Time Series. First we apply our analysis of labor demand to a time series of U.S. copyright. Here we must assume that the distribution $M(\rho)$ is time invariant, and 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,[13] and the amount of labor in the sector by the number of copyright registrations. The relevant annual growth rates for the U.S. are reported, by decade, in Figure 5.1.

[13]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.

FIGURE 5.1. _



If elasticity of total monopoly revenue is constant or decreasing, we expect to see per capita copyright growing more rapidly than population. This is in fact the case prior to 1900 and after 1970, but those are both anomalous periods. 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.[14] Similarly, in 1972 it became possible to copyright musical recordings other than phono 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 surprising that copyright registrations jumped up 1972. In 1976, the term of copyright, which since 1909 had been 28 years, plus a renewal term of 28 years, 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 1920-1930 and 1960-1970 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 entirely new areas such as movies, recorded

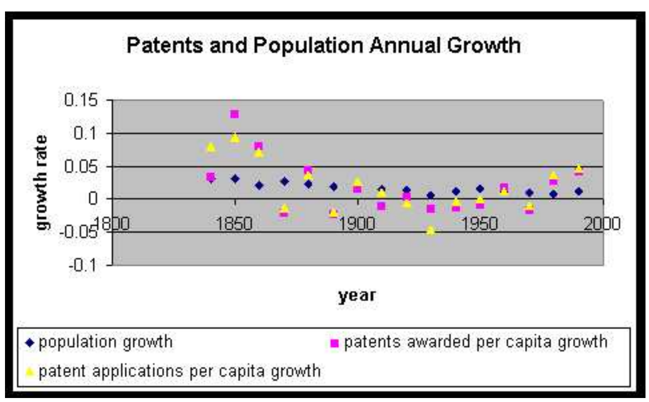
[14] A brief history of U.S. copyright can be found at U.S. Copyright Office [2001a]. The 1972 change is described in U.S. Copyright Office [2001b].

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 U.S.A. *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 U.S.A., the second because the growth of per capita income and, especially, the expansion of "American culture" around the world greatly increased the potential market size.

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 non-idea 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 non-idea 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 5.2 is the patent analog of Figure 5.1 and is quite similar.

FIGURE 5.2. _



Whether we measure patentable activity by patents awarded or by 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. In summary, the time series of patents lead us to the same conclusions we reached with copyright: that patents have grown less than market size, thereby suggesting that the elasticity of monopoly revenue is increasing also in this case.

An alternative to measuring either patent applications or awards is to use R&D expenditure as a proxy for the amount of labor used in creating new ideas. 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 fairly fuzzy and available only for relatively recent years – the major source of data being an NSF 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 ambiguous and yet different from that of the number of patents – ambiguous because the choice of *which measure* of R&D expenditure one should consider is not obvious. One possibility is to focus on the private sector only. However, we would expect that research financed by the federal government – much of which is carried out at private institutions – both produces useful ideas and increases the demand for skilled labor. On the other hand, there are reasons to believe that the federal expenditure in R&D reacts much less, or maybe not at all, to market incentives and to the expected profitability of innovations in particular.[15] Universities, either public or private, are obviously producing ideas and employing skilled workers, but the extent to which they respond to market incentives may have varied substantially during the last fifty years. In the light of this, we will report statistics for four

[15] This point is made by Jones [2004] while analyzing the R&D data and the "patent puzzle." However, we would expect some scale of market effect on federal R&D expenditure as well – as the scale of the market increases so does the tax base that pays for the expenditures.

aggregates: total, private sector plus universities, and these same two series adjusted for the wage rate of college and post-college workers. The latter are relevant because the wage skill premium increased dramatically during the last thirty years, and workers involved in R&D activities hold college, and most often post-college, degrees.

The ratio of total R&D expenditure to GDP has grown from 1.36% in 1953 to 2.78% in 2002, thereby doubling in fifty years. During the same time, population has grown about 80% and real GDP has almost quintupled. It may be worth noticing that the maximum value for the total R&D expenditure to GDP ratio, 2.88%, was reached in 1964. For the private plus universities aggregate, the same ratio has more than tripled between 1953 and 2002, going from 0.63% to 2.0%. Next, assume that the cost of labor employed in the idea sector grows, roughly, at one-half the college wage and one-half the post-college wage.[16] Then the cost of the average worker in the idea sector between 1963 and 2002, the period for which data are available, has grown by about 95%, while over the same period, the mean wage has grown by about 65%.[17] Between 1963 and 2002, the ratio of total R&D expenditure to GDP basically does not move, while the industry plus universities ratio goes from 0.9% to 2.0%. That is, the industry plus university ratio grows by 110%, population grows by 52%, and total GDP by 70%. Because our index of the relative wages in the idea sector has grown roughly 20% over the same period, it turns out that, if one uses total expenditure in R&D, then the share of workers in the idea sector has actually declined, implying a strongly increasing elasticity of $M(\rho)$; if, instead, one uses the private plus universities measure, it has grown by about 90%. The latter is somewhat higher than either the population or the GDP growth rates; hence, on the basis of the last index, one cannot rule out the hypothesis that the elasticity of the total revenue function is either constant or decreasing.[18]

R&D Cross Section. Finally, we look at a cross section of countries. Here we run a simple cross-country regression with R&D as a fraction of GDP

[16]This is arbitrary but not unreasonable.

[17]High school graduate wages grew 20%, college graduate wages grew by 65%, and post-college graduate wages grew at 123%; see Eckstein and Nagypal [2004], Figures 1 and 3.

[18]The number of additional caveats one would need to add is endless. The tax and accounting treatments of R&D have both changed substantially over the period, both favoring the rellabeling of many sources of cost as R&D expenditure. The Cold War and the often changing policies of the federal government with respect to financing basic research, carrying it out directly, or privatizing through subsidies also add additional uncertainty to the interpretation of the data.

as the dependent variable and market size and the strength of IP protection as explanatory variables.[19] 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 of labor demand with respect to market size, we can write $\log \ell = \vartheta \log(\phi\lambda)$. To account for the effect of both population and per capita GDP on market size, write $\lambda = y^{\alpha}N$, where N is population and y is per capita GDP.[20] Ordinary least squares regression gives $\vartheta = 0.20(0.03)$ and $\alpha\vartheta = .56(0.038)$, meaning that $\alpha = 2.8$, a remarkably large number that, if applied to the previous time series analysis would imply a strongly increasing elasticity.[21] The estimated 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 were on the order of 5, it would be possible for the elasticity of per capita R&D with respect to size of market to be near unitary. However, a ratio of 5 is implausibly large for most countries with the exception, possibly, of Switzerland and Luxembourg. Exports are almost everywhere a fraction, not a multiple of GDP. Consequently, a ratio of 5 would be possible only if R&D were much more intensive in export industries than the average – by a factor larger 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 micro-data on 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 U.K. and Germany, with most countries receiving 15-20% of income from a patent from rights held abroad. So the evidence easily contradicts the idea that $\lambda_{world}/\lambda_{average}$ is on the order of 5.

[19] To measure the latter we use an index developed by Walter Park, to whom we are grateful for providing us with his data. Details of the construction can be found in Park and Lippholdt [2003].

[20] Standard errors in parentheses, $R^2 = 0.65$.

[21] The underlying data include 34 countries for the period 1980-1997, and can be found at http://www.dklevine.com/data.htm.

6. Implications for National IP Policy

What consequences does our analysis have for the optimal IP policy? The first set of calculations indicates that IP protection for patents is probably too high, but this conclusion is somewhat tentative. In the case of copyright, it seems conclusive that copyright terms are far too long. The second set of calculations strongly indicates that the scale of market effect is quantitatively significant and that there should be substantial reductions in the length of IP term in response to size of market increases.

The first step is to translate ϕ – the effective degree of protection – into the relevant policy parameter - the length of term. This depends on the interest rate and on depreciation.

Length of Term, Depreciation, and Effective Protection. Suppose that the real interest rate is r, that all ideas depreciate at a common rate d and that the length of term is T. Then – with perfect enforcement – the effective protection is $\phi = 1 - e^{-(r+d)T}$. Reasonable estimates of the real interest rate lie between 2% and 4%. Since the Sony Bono Copyright Term Extension Act of 1998, copyright protection in the U.S. 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.[22] Current patent length in the U.S. for utility patents (inventions) is 20 years.

Depreciation rates are more difficult. In our book 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 over about one-third of a year, or an annual depreciation rate of nearly 95%.[23] Capital goods depreciation rates are generally thought to be close to 8% per year, including housing and building, which depreciate more slowly. Little data are available about depreciation rates of idea so insofar as ideas correspond to generations of capital, they may well depreciate at the same rate; some very good ideas (the law of gravity) may not depreciate at all.

If the flow of sales is constant over time, for a copyright length of $T = 105$ years, and different interest rates r and depreciation rates d, the corresponding values of $\phi = 1 - e^{-(r+d)T}$ are given in Figure 6.1.

---

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

[23] This is consistent with data for the other months and with the general claim that the most significant book sales occur within three months of publication.

FIGURE 6.1. Effective Copyright Protection

$r+d$	$T=20$	$T=105$
0.02	0.330	0.878
0.03	0.451	0.957
0.04	0.551	0.995
0.07	0.753	0.999
0.08	0.798	1.000
0.09	0.835	1.000
0.38	1.000	1.000

The low values 0.02, 0.03, 0.04 for $r+d$ correspond to no depreciation; the intermediate values 0.07, 0.08, 0.09 correspond to a modest depreciation rate of 5%; we do not report any values larger than 0.38 (that is, depreciation between 34% and 36%) since, even with just a 20-year term, $\phi = 1$ at this point. In summary, for realistic interest and depreciation rates, the current copyright term certainly corresponds to $\phi = 1$ in our model, while current patent terms correspond to roughly $\phi = 0.9$.

The Static Optimum. To determine the optimal level of protection we can solve the NOC to find

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

To proceed we need to know $\overline{\nu}^C$ and $\overline{\nu}^M$, beside Υ. We will consider the benchmark case of linear demand and constant marginal cost so that $\overline{\nu}^C = 2$ and $\overline{\nu}^M = 3/2$. Below we examine alternative demand structures and argue that this benchmark case is empirically relevant.

In Figure 6.2 we report (second column) the optimal values of $\hat{\phi}$ corresponding to elasticities in the empirically relevant range (first column). The other two columns translate the optimal $\hat{\phi}$ in lengths of term, using different interest and depreciation rates.

FIGURE 6.2. Optimal Protection and Term Length

Υ	$\hat{\phi}$	$r+d=0.2$	$r+d=0.4$	$r+d=0.08$
0.03	0.13	7	4	2
0.10	0.24	14	7	4
0.15	0.33	20	10	5
0.20	0.40	26	13	7
0.30	0.51	36	18	9
0.40	0.60	46	23	12

Two facts stand out. First, optimal length of protection is less than 1 – meaning that given that elasticity is increasing, optimal copyright and patent protection should decline with the size of the market. Second, in the case of copyright, 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 this data; if we consider depreciation – not in the empirical range of 95%, but say in the range of 5% – copyright protection should be at most several years. This is generally consistent with our scale of market calculations below under the hypothesis that 28 years at the turn of the 20th century was about right.

In the case of patents, estimated elasticities appeared somewhat larger, with .15 being a sensible middle ground estimate. With a real interest rate plus depreciation rate of 4%, this implies an optimal patent length of 10 years, while with a more realistic depreciation adjustment it would be closer to 5 years - again, not so terribly different than what we would get if we assumed term length were correct at the beginning of the twentieth century. If we took the high end elasticity of .4 and a real interest rate of just 2%, the optimal term would be 46 years; hence it is not impossible, at least in principle, to reconcile existing patent term with available data. Realistic estimates, though, suggest that optimal patent term should be between 5 and 10 years.

The Scale of Market Effect. To examine the scale of market effect, we differentiate the NOC to find

$$\mathfrak{E}\phi(\lambda) = -\frac{\lambda}{\phi}\frac{NOC_\lambda}{NOC_\phi} = -\frac{1}{1+(1+\Upsilon)/\mathfrak{E}\Upsilon+(1/(\overline{\nu}^C-\overline{\nu}^M))}.$$

To get a feeling for this, note that in the simple and empirically relevant case that $M(\rho)$ is linear $\mathfrak{E}\Upsilon = 1+\Upsilon$. Consequently $\mathfrak{E}\phi(\lambda)$ is $-1/2$ or less negative depending on $\overline{\nu}^C-\overline{\nu}^M$. When demand is linear $\overline{\nu}^C-\overline{\nu}^M = 1/2$, and $\mathfrak{E}\phi(\lambda) = -1/4$. This means that a 10% increase in size of market should reduce effective protection by 2.5%. For example, if the world economy is growing at 4% per year, then a simple rule of thumb would be to reduce protection by about 1% per year. In the case of 20-year patents that would mean about two months each year. One implication of this is that during the last century in which world GDP grew by a factor of roughly 40, optimal protection should have declined from 20 years to about 1 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 "golden records," while in the current times a successful record sells easily ten or twenty million copies. The effective size of the market has, therefore, increased 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 twelve. 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 a hundred 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 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 20th century was socially optimal, the current term should be a little over a year, rather than the current term of approximately 100 years. This gives a ratio of 100 between the actual copyright terms and their socially optimal value.

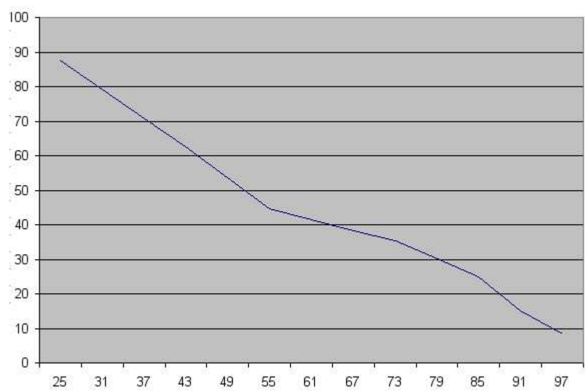
Product Demand. The linear demand benchmark is particularly useful, because it implies return neutrality, and, provided marginal cost is constant, it is independent of it. But is linear demand empirically relevant?

Take first the case of a small cost-saving innovation – for example, a way of making a machine work a little better. This is the type of thing most people think of when they think of an "invention," although only a small fraction of patents are of this type. Demand for a small cost-saving innovation is equal to the per machine cost saved up to the number of machines – then drops to zero. Since the innovation is small it has an insignificant effect on the number of machines. In this case to a good approximation $\overline{v}^C = \overline{v}^M = 1$, since we have normalized so that the monopoly profit is 1. In this case the elasticity of total monopoly revenue does not really matter: the social optimum is to set $\phi = 1$, and it does not change in response to the scale of market.

More generally, it is easy to see that if demand is concave, then $\overline{v}^M, \overline{v}^C - \overline{v}^M$ are smaller than in the linear case – the extreme case being that of a small cost-saving innovation – while if demand is convex then $\overline{v}^M, \overline{v}^C - \overline{v}^M$ are larger than in the linear case. Notice that larger $\overline{v}^M$ and $\overline{v}^C - \overline{v}^M$ increase the scale of market effect, but have an ambiguous effect on the level of IP: larger $\overline{v}^M$ tending to increase and larger $\overline{v}^C - \overline{v}^M$ tending to decrease optimal IP.

In understanding how good the benchmark linear case is, it is important to recognize that demand for most innovations is strongly affected by income. Take the case of new drugs: it is probably a good approximation to think of willingness to pay as proportional to individual income. From 2001 census data for the U.S., assuming that each individual demands one unit of an innovation, with willingness to pay proportional to income, we construct the demand curve shown in Figure 6.3.

FIGURE 6.3. Demand Proportional to Income



U.S. Income Distribution 2001
Source: U.S. Census

In other words, to a good approximation, demand based on linear Engel's curves is to a good approximation linear. Artistic creations such as books, movies, and music are similar to drugs in that demand is heavily dependent on income. In fact drugs and artistic creations are undoubtedly superior goods, meaning that the fraction of income spent on them increases as income goes up.

If we start with linear demand for goods that have linear Engel's curves, then goods that are strongly superior, in the sense that the increased fraction of income spent rises at an increasing rate, have convex demand curves. Conversely for goods that are strongly inferior – orphan drugs are a likely example – demand will be concave.

The conclusion is that for most types of goods, the linear demand approximation is conservative – most likely overstating the level of optimal IP protection and understating the optimal rate of decrease in response to market size. The exception is in the case of small cost-saving innovations – which to a certain extent matches the idea in patent law of "process" rather than "product" patents. Historically "process" patents – patents on methods for doing things – have received stronger protection than "product" patents.

The theory indicates that this is in fact the right approach. Unfortunately, despite the great historical success – for example, in the development of the chemical industry – of allowing only "process" patents in countries such as Germany, the Anglo-French system of allowing products the same protection as processes has become widespread.

One final consideration is that the values of $\overline{v}^C$ and $\overline{v}^M$ are measured relative to private profit, and so depend on the scope of copyright or patent. If the scope is relatively narrow, then for a given social value of the patent, a lower private return can be obtained, due to competition. If we let π be the amount of profit lost to competition, then $\overline{v}^C$, $\overline{v}^M$ and their difference are all multiplied by $1/(1-\pi)$. We see, for example, that narrower scope implies higher optimal effective protection – but also a more rapid decline with the scale of the market.

7. Implications for Trade and Harmonization Policies

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.

It is tempting to view this as a typical tariff-like free-riding problem: countries try to free ride off 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.[24]

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

[24] Grossman and Lai [2004] correctly identify the first effect, but due to an algebraic error, miss the second. In their 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.

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, which 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 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,

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

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, $\varepsilon_i = 1$ 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.

Given everyone else's choice of protection, $\phi_j \;\; j \neq i$, the optimum of an individual country i is calculated by differentiating the social welfare function with respect to ϕ_i to get

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

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 $NOC(\phi_i, \phi) = 0$.

Our basic model then follows Grossman and Lai [2004] in assuming a noncooperative game in which each country maximizes their own social welfare without harmonization, and a cooperative game in which the common policy is chosen to be the worldwide optimum with harmonization. Of course there is also the possibility that the bargaining under harmonization will not lead to a worldwide optimum. This issue is discussed in length in Scotchmer [2004] who argues that bargaining is likely to lead to excessive IP protection at the expense of public sponsorship of research.

The Symmetric Case. Consider first a symmetric equilibrium of a symmetric model in which $\theta_i = \varepsilon_i = 1/I$. Holding λ constant, let ϕ^1 be the solution to the single country problem, and ϕ^I the symmetric solution to the multi-country NOC above. Because the first term of the NOC is positive and the second negative ($I > 1$), and because the second terms 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 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 under protect 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, 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 Figure 7.1.

FIGURE 7.1. GDP and Invention in the South

	GDP Trillion U.S.$	% of World GDP (θ_i)	% of U.S. 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 $\varepsilon_1 = .987$ 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. In the asymmetric case, smaller countries have an incentive to free ride off the larger countries. First we show as a matter of theory that, regardless of I, the equilibrium level of aggregate protection, ϕ, is bounded away from zero, and small countries have little incentive to adopt IP protection. There is one large country with shares θ_1, ε_1 and $I-1$ small countries with shares $\theta_i = (1-\theta_1)/(I-1) < \theta_1$, $\varepsilon_i = (1-\varepsilon_1)/(I-1) < \varepsilon_1$. We prove the following proposition in Appendix 1.

Proposition 7.1. *Let $\widetilde{\phi}^1$ be the unique solution to $NOC(\widetilde{\phi}^1, \theta_1\widetilde{\phi}^1) = 0$. Then in any equilibrium $\phi^1 \geq \widetilde{\phi}^1 > 0$ and $\phi \geq \theta_1\widetilde{\phi}^1 > 0$. If the number of small countries is bigger than*

$$I^* = \frac{(\overline{\nu}^C - \overline{\nu}^M + 1)}{(1-\theta_1)\left[\frac{\overline{\nu}^C}{\theta_1\widetilde{\phi}^1} - 1\right]\Upsilon(1/\theta_1\widetilde{\phi}^1\lambda) + \varepsilon_1} + 1$$

the equilibrium level of protection for a small country, $\phi_i = 0$.

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

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

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 first that $\varepsilon_i = \theta_i$ 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 $\phi_1 \geq 0.3$ 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, even remotely, 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.

Harmonization, then, requires an increase in IP in the South. Grossman and Lai [2004] show that in addition the overall level of IP must increase throughout the world, so the increase in the South must more than offset any decrease in the North. However, this leaves open the question of whether IP in the North goes up or down.

Why There Is Too Much IP in the North. Consider first the case in which $\varepsilon_1 = 1$ and $\theta_1 < 1$, that is, all ideas are produced in the North. In this case the solution for the North 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 market size 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 $\varepsilon_1 < 1$, this effect is weakened. This is because of the "profit stealing" effect mentioned above. When the South is not protecting at all, the North 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 South also produce ideas, then part of the increased royalty payments made by consumers in the North is lost to Southern monopolists. Since in this case the North sets a lower level of IP than when $\varepsilon_1 = 1$, 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]. It is then crucial to check empirically which effect dominates in the "size of market versus profit stealing" game, which is what we do next.

Assume, then, that $\phi_i = 0$ for all small countries (the South), the NOC for the North is

$$2\left[1+\frac{1-\phi_1}{\phi_1}\right]\Upsilon\left(\frac{1}{\theta_1\phi_1\lambda}\right)=\frac{3}{2}-\varepsilon_1,$$

while the NOC for the harmonized welfare maximum is

$$2\left[\frac{1}{4}+\frac{1-\phi}{\phi}\right]\Upsilon\left(\frac{1}{\phi\lambda}\right)=\frac{3}{2}-\varepsilon_1,$$

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

We take $\theta_1 = .75$ throughout, but even when $\theta_1 = .66$ (that is, the North is just the G7), the conclusion holds. The RHS of the North's NOC is $3-2\varepsilon_1$ times the RHS for the harmonized welfare maximum. On the other hand, if world protection under harmonization were the same as the one chosen by the North when going alone, then the LHS of the North's NOC is the elasticity of Υ times $1/\theta_1 = 4/3$. In the case in which $M(\rho)$ is linear, the elasticity $\rho\Upsilon'/\Upsilon = 1+\Upsilon$, so the LHS of the North's NOC is $(4/3)(1+\Upsilon)$ times the LHS of the harmonized NOC. By the second order condition, it follows that it is optimal for the North to set its protection above the harmonized social optimum if and only if $(1+\Upsilon) > (9/4)-3\varepsilon_1/2$. What matters, then, are the actual magnitudes of ε_1 and of the elasticity of the total revenue function.

Suppose first that the South is just as effective at producing ideas as the North so that $\varepsilon_i = \theta_i = .75$. In this case we see that the large country should reduce protection upon harmonization if $\Upsilon \geq .125$, which is certainly the case for our range of estimates of Υ.

However, the assumption that $\varepsilon_i = \theta_i$ is not nearly true. As Figure 7.1 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 $\varepsilon_1 = 0.987$. With this calibration $1/\theta_1 > 3-2\varepsilon_1$ so that the North is unambiguously setting its level of IP protection too high. By going to the data, we have not only learned that, given reasonable parameter values, the model predicts what we seem to observe – that is, that the North is overprotecting – but also found an answer to the question of "why" it does so. It is not because the planner is maximizing the utility from ideas of its $\theta_1 = .75$ "effective consumers" but because it is maximizing the profits of its $\varepsilon_1 = .987$ "IP monopolists." Clearly, upon harmonization the North should reduce protection regardless of the elasticity Υ; the open question is, by how much?

By How Much Should the North Decrease IP under Harmonization? 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 North with that for the harmonized welfare maximum, that is, the two values of ϕ that solve our two NOC above. Let us use the superscript 1, as in ϕ^1, to denote the elements of the North's NOC, with no superscript for the other NOC. Then we can solve either of the NOCs to compute the associated optimal level of IP protection. So, for example, the solution for the North's NOC yields

$$\phi^1 = \frac{4}{3 + 2RHS^1/\Upsilon^1},$$

with an analogous formula for the other case. Dividing the two solutions and using the approximation that $\Upsilon^1/\Upsilon \approx (1+\Upsilon^1)(\phi/\phi^1\theta_1)$, we get

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

The resulting quadratic solves as

$$\frac{\phi}{\phi^1} = \frac{-3\Upsilon^1 + \sqrt{9\left(\Upsilon^1\right)^2 + 4\left(3\Upsilon^1 + (3 - 2\varepsilon_1)\right)\left(1+\Upsilon^1\right)/\theta_1}}{2\left(1+\Upsilon^1\right)/\theta_1}.$$

For Υ in the range 0.15-0.40 this implies ϕ^1/ϕ in the range 0.817 to 0.845, given our maintained calibration for θ_1 and ε_1. 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, 20 years to 16 years.

We have treated the North as a unitary actor. If the U.S., Europe and Japan are not themselves harmonized, then they will tend to underprotect, as in the symmetric case, and it may be that the optimum is an increase in the North. However the evidence that the North is internally harmonized seems to us quite strong; indeed the U.K. is currently debating further extending copyright on existing works in order to better harmonize with the U.S.

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 it that the social planner in the North is not setting current copyright levels to maximize social welfare in the North but, we conjecture, to maximize the rents accruing to the interest groups that are covered by the copyright laws. We cannot pursue this line of investigation here, but the theoretical and empirical evidence we have collected in this paper suggests that a positive theory of IP protection based on the rent-seeking explanation deserves, at least, a shot.

8. CONCLUSION

Our goal in this paper has been to reconcile with the economic theory of optimal IP protection, a variety of data from different sources, ranging from book revenues to patent values estimated by renewal rates, to R&D expenditures. In the case of copyright, we think that evidence in favor of increasing elasticity of total monopoly revenue is decisive and that existing copyright terms are vastly too long: 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, but the best available evidence suggests 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. And our quantitative analysis indicates that the scale of market effect is strong and that as a consequence there should be both an immediate reduction in patent terms as part of any TRIP agreement augmented by a phased reduction tied to future growth.

Our theoretical results are robust in a variety of dimensions, while our methods indicate directions along which data collection can be improved. On the theoretical side, we assumed that utility is linear in the output of the idea sector and in labor. We can consider more generally the possibility that utility is concave in those variables. This strengthens the scale of market effect: As the market grows and more ideas are produced, the price in the idea sector declines and the cost of labor increases because more labor moves to the idea sector. Therefore, it is still best to exploit the opportunity offered by an increase in the size of the market by reducing protection, rather than by allowing the relative price of skilled labor to rise.

If, following Boldrin and Levine [1999, 2002, 2004, 2005], we recognized that, absent any IP protection, profits for innovators would still be substantial, the scale of market effect would also be strengthened. Indeed, it can easily be shown that the optimal level of protection should be set equal to zero at a finite market size, not just asymptotically as in the extreme case of no competitive rents. This also highlights the greatest theoretical deficiency of standard models – the lack of adequate dynamic analysis. A particular problem is the fact that ideas build on other ideas. As pointed out in Scotchmer [1991] and Boldrin and Levine [1999, 2004, 2005], 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. Indeed, when the complexity of innovations increases because new ones need to use more and more old ideas as inputs, the presence of widespread IP creates a hold-up problem where even one residual monopolist may prevent new ideas from being implemented as in Boldrin and Levine [2005].

We have largely ignored also the important issue of rent-seeking. Suppose, for example, that the size of the indivisibility does not vary systematically with private returns. Then ideas with high returns are also more likely to have high absolute levels of profit associated with them. If it is possible to purchase "extensions" of protection 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 extension 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.

Uneven depreciation of ideas has a similar effect. The basic model assumes that ideas depreciate at the same rate – that is, a given real time length of protection implies the same ϕ for all ideas. There is some reason to think that better ideas depreciate less quickly than the average. Leibowitz and Margolis [2003] in particular argue that good books generate a substantial revenue stream for a much longer period of time than poor books. Certainly ideas that generate revenue for longer periods are, *ceteris paribus*, better than those that do not, so there is reason to suspect a positive correlation. Suppose, then, that good ideas depreciate more slowly than bad ideas, so a high ρ means also a longer economic life. Extending the length of IP protection does little to encourage the production of marginal ideas in this case, since these depreciate very rapidly anyway. It simply serves to increase rents and the monopoly distortion on good ideas. In the extreme case of a one-horse-shay depreciation – ideas last a fixed length of time then collapse – there is never any reason to set the length of protection longer than the duration of the marginal idea.

REFERENCES

[1] Acemoglu, D. and J. Linn (2003), "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," mimeo, MIT.

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

[3] 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 Court of the United States, Eldred versus Ashcroft*.

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

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

[6] Boldrin, M. and D.K. Levine (2003), "Rent Seeking and Innovation," *Journal of Monetary Economics*, **51**, 29–41.

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

[8] Boldrin, M. and D.K. Levine (2005), "The Economics of Ideas and Intellectual Property," *Proceedings of the National Academy of Sciences*, **102**, 1252–1256.

[9] DiMasi, J. A., R. W. Hansen, H. G. Grabowski, L. Lasagna (1991), "The Cost of Innovation in the Pharmaceutical Industry,*" Journal of Health Economics,* **10,** 107–142.

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

[11] 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.

[12] Gallini, N. (1992), "Patent Policy and Costly Imitation," *Rand Journal*, **23,** 52-63.

[13] Gilbert, R. and C. Shapiro (1990), "Optimal Patent Length and Breadth," *Rand Journal*, **21,** 106-112.

[14] Grossman, G. M. and E. Helpman (1991), "Trade, Knowledge Spillovers and Growth," *European Economic Review (Papers and Proceedings)*, **35,** 517–526.

[15] Grossman, G. M. and E. Helpman (1994), "Endogenous Innovation in the Theory of Growth," *Journal of Economic Perspectives*, **8,** 23–44.

[16] 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.

[17] Grossman, G. M. and E. L. Lai (2004), "International Protection of Intellectual Property," *American Economic Review* **94**, 1635–1653.

[18] Hall, B. (2001), "NBER Patents Data File," http://elsa.berkeley.edu/users/bhhall/pat/datadesc.html.

[19] Hall, B., A. Jaffe and M. Tratjenberg (2004), "Market Value and Patents Citations," *Rand Journal of Economics*, forthcoming.

[20] 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.

[21] Jones, C. (2004), "Growth and Ideas," prepared for the *Handbook of Economic Growth*.

[22] Kanwar, S. and R. E. Evanson (2001), "Does Intellectual Property Protection Spur Technological Change?" Cowles.

[23] Landes, W. M. and R. A. Posner (2003), *The Economic Structure of Intellectual Property Law*, Harvard University Press.
[24] Lanjouw, J. (1993), "Patent Protection: Of What Value and How Long?" NBER Working Paper 4475.
[25] 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.
[26] Leibowitz, S. and S. Margolis (2003), "Seventeen Famous Economists Weigh in on Copyright: The Role of Theory, Empirics, and Network Effects," University of Texas at Dallas.
[27] Lo, S. (2003), "Strengthening Intellectual Property Rights: Experience from the 1986 Taiwanese Patent Reforms," UCLA.
[28] Maurer, S. M. and S. Scotchmer (2002), "The Independent Invention Defense in Intellectual Property, *Economica,* forthcoming.
[29] Pakes, A. (1986), "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, **54**, 755–784.
[30] 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.
[31] Romer, P. M. (1990), "Endogenous Technological Change," *Journal of Political Economy*, **98**, s71-s102.
[32] Romer, P. (1994), "New Goods, Old Theories, and the Welfare Costs of Trade Restrictions," *Journal of Development Economics*, **43**, 5-38.
[33] Scotchmer, S. (1991), "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law," *Journal of Economic Perspectives,* **5,** 29–41.
[34] Scotchmer, S. (1999), "On the Optimality of the Patent Renewal System," *Rand Journal of Economics*, **30**, 181–196.
[35] Scotchmer, S. (2004), "The Political Economy of Intellectual Property Treaties," *Journal of Law, Economics and Organizations*, **20**, 415–437.
[36] 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.
[37] Taylor, M. S. (1993), "TRIPS, Trade, and Technology Transfer," *Canadian Journal of Economics,* **26**, 625–637.
[38] Taylor, M. S. (1994), "TRIPS, trade, and growth," *International Economic Review,* **35**, 361-381.
[39] U.S. Copyright Office (2001a), *A Brief History and Overview*, Circular 1a, http://www.copyright.gov/circs/circ1a.html.
[40] U.S. Copyright Office (2001b), *Copyright Registration for Sound Recordings*, Circular 56, http://www.copyright.gov/circs/circ56.html.
[41] 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 1: Proofs

Proposition. 3.1. *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* $\lambda > \overline{\lambda}$*. Further, when* $\hat{\phi}(\lambda) < 1$*, in a neighborhood of* $\rho = 1/\lambda\hat{\phi}(\lambda)$*, the following three cases hold. (I)* $\Upsilon'(\rho) > 0$ *implies* $\hat{\phi}(\lambda)$ *is unique and strictly decreasing; (II)* $\Upsilon'(\rho) = 0$ *implies* $\hat{\phi}(\lambda)$ *is unique and locally constant; and (III)* $\Upsilon^{\rho} < 0$ *and* $\hat{\phi}(\lambda)$ *unique*[25] *implies* $\hat{\phi}(\lambda)$ *is strictly increasing.*

Proof. Use return neutrality to rewrite social welfare as

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

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

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

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

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

has the same qualitative properties as $FOC(\lambda,\phi)$: it has the same zeros, 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

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

The second term is unambiguously negative. The first term has two factors of interest. We have $(1/\phi)\left\{\phi\overline{v}^M + (1-\phi)\overline{v}^C\right\} - 1$ representing social surplus of the marginal idea produced; since privately it yields zero profit, it must yield positive social surplus. If the other factor $\Upsilon'(1/\lambda\phi) >$

[25] In this case we cannot guarantee that the second order condition is satisfied, so we must rule out the possibility that $\hat{\phi}(\lambda)$ has multiple values.

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

$$\frac{d\phi}{d\lambda} = -\frac{NOC_\lambda}{NOC_\phi} \propto NOC_\lambda =$$

$$= -\left[(1/\phi)\left\{\phi\overline{v}^M + (1-\phi)\overline{v}^C\right\} - 1\right](1/\lambda^2\phi)\Upsilon'(1/\lambda\phi),$$

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 small enough ρ we may assume that either $\Upsilon'(\rho) > 0$ or $\Upsilon'(\rho) < 0$. 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 non-increasing. 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 $\Upsilon'(\rho) > 0$ near $\rho = 0$, so fix $\phi = 1$ and examine

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

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 $D_\lambda L^D \geq L + \varepsilon$ 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, $D_\lambda L^D \to \infty$, so in this case the labor constraint must bind. □

Proposition. 3.2. *Consider two different 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, $\mathfrak{E}[\ell_1(\rho)] > \mathfrak{E}[\ell_2(\rho)]$.*
(3) *As the elasticity of total revenue goes from increasing, to constant, to decreasing, the elasticity of the associated labor demand functions increases 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
$$D\Upsilon(\rho) = D[-\rho DM(\rho)/M(\rho)] =$$
$$= \frac{1}{\rho}[\Upsilon(\rho) + \Upsilon^2(\rho)] - \rho D^2M(\rho)/M(\rho)$$
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: $\mathfrak{E}[\ell_1(\rho)] > \mathfrak{E}[\ell_2(\rho)]$
Because

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

$\mathfrak{E}[\ell_1(\rho)]$ and $\mathfrak{E}[\ell_2(\rho)]$ have the same numerator, and, because of Step 2, the first has a smaller denominator. Hence the conclusion. □

Proposition. 7.10. *Let* $\widetilde{\phi}_1$ *be the unique solution to* $NOC(\widetilde{\phi}_1, \theta_1\widetilde{\phi}_1) = 0$. *Then in any equilibrium* $\phi_1 \geq \widetilde{\phi}_1 > 0$ *and* $\phi \geq \theta_1\widetilde{\phi}_1 > 0$. *If the number of small countries is bigger than*

$$I^* = \frac{(\overline{\nu}^C - \overline{\nu}^M + 1)}{(1-\theta_1)\left[(1/\theta_1\widetilde{\phi}_1)\overline{\nu}^C - 1\right]\Upsilon(1/\theta_1\widetilde{\phi}_1\lambda) + \varepsilon_1} + 1$$

the equilibrium level of protection for a small country, $\phi_i = 0$.

Proof. The NOC for the large country is

$$NOC_1(\phi_1, \phi) = \frac{\theta_1}{\phi}\left[\phi_1\left(\overline{\nu}^M - 1\right) + (1-\phi_1)\overline{\nu}^C\right]\Upsilon(1/\phi\lambda) - (\overline{\nu}^C - \overline{\nu}^M + 1 - \varepsilon_1) = 0.$$

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

$$NOC_i(0, \phi) = \frac{(1-\theta_1)}{(I-1)}\left[\frac{1}{\phi}\overline{\nu}^C - 1\right]\Upsilon(1/\phi\lambda) - (\overline{\nu}^C - \overline{\nu}^M + 1 - \frac{\varepsilon_1}{I-1})$$
$$\leq \frac{(1-\theta_1)}{(I-1)}\left[\frac{1}{\widetilde{\phi}_1}\overline{\nu}^C - 1\right]\Upsilon(1/\theta_1\widetilde{\phi}_1\lambda) - (\overline{\nu}^C - \overline{\nu}^M + 1 - \frac{\varepsilon_1}{I-1})$$

which is strictly negative for I larger than

$$I^* = \frac{(\overline{\nu}^C - \overline{\nu}^M + 1)}{(1-\theta_1)\left[(1/\theta_1\widetilde{\phi}_1)\overline{\nu}^C - 1\right]\Upsilon(1/\theta_1\widetilde{\phi}_1\lambda) + \varepsilon_1} + 1.$$

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

APPENDIX 2: DATA

Book Revenue. We collected all the titles, ISBN numbers, and sale prices listed by 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 number 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 on www.amazon.com. Ingram is a large book distributor, and generally thought to generate roughly one-sixth of all book sales. It should be noted that the sales prices on 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 it is available from http://www.dklevine.com/data.htm.

Copyright Time Series. The basic source of the copyright registration time series is from 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 1901-1999 is from the U.S. Census http://www.census.gov/population/estimates/nation/popclockest.txt; data prior to 1901 is from http://www.census.gov/population/censusdata/table-2.pdf; the two sources have a slight discrepancy for the 1900 population with the former source reporting 76,094,000 and the latter (which we used) 76,212,168. The year 2000 data was 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 U.S., 2004-2005.