PERFECTLY COMPETITIVE INNOVATION

MICHELE BOLDRIN AND DAVID K. LEVINE

ABSTRACT. We construct a competitive model of innovation and growth under constantreturns to scale. Previous models of growth under constant returns cannot model technological innovation. Current models of endogenous innovation rely on the interplay between increasing returns and monopolistic markets. We argue that ideas have value only insofar as they are embodied in goods or people, and that there is no economic justification for the common assumption that ideas are transmitted through costless "spillovers." In the absence of unpriced spillovers, we argue that competitive equilibrium without copyrights and patents fails to attain the first best only because ideas are indivisible, not because of increasing returns. Moreover, while it may be that indivisibility results in socially valuable ideas failing to be produced, when new ideas are built on old ideas, government grants of intellectual monopoly may lead to even less innovation than under competition. The theory of the competitive provision of innovations we build is important both for understanding why in many current and historical markets there has been thriving innovation in the absence of copyrights and patents, and also for understanding why, in the presence of the rent-seeking behavior induced by government grants of monopoly, intellectual property in the form of copyrights and patents may be socially undesirable.

Date: First Version: October 3, 1997, This Version: Jan 17, 2003.

Many ideas presented here appeared first in an unpublished (1997) paper: "Growth Under Perfect Competition." Both authors thank the National Science Foundation and Boldrin thanks the University of Minnesota Grants in Aid Program for financial support. We benefited from comments from seminar participants at Toulouse, the London School of Economics, Humboldt University, UC Berkeley, Cornell, Chicago, Wisconsin-Madison, Iowa State, New York University, Stanford, Univ. of Pennsylvania, Columbia, Oxford, CEMFI, Carlos III, Rochester, DELTA-ENS Paris, Venice, Padova, IGIER-Bocconi, and the University of Minnesota. Jim Schmitz also made a number of valuable suggestions.

1. INTRODUCTION

This paper is about technological change, defined as the invention and subsequent adoption of new goods and techniques of production. As much of the history of technological change has taken place in the absence of copyrights, patents and other forms of legal monopoly, it is important to understand how innovations take place in a competitive environment. We argue that unpriced externalities are probably not an important element of this process. We argue, instead, that ideas have value only insofar as they are embodied in goods or people, and that there is no economic justification for the common assumption that ideas are transmitted through costless "spillovers." In the absence of unpriced spillovers, we show that competitive equilibrium without copyrights and patents may fail to attain the first best only because ideas are indivisible, and such indivisibility may occasionaly bind, not because of increasing returns. Moreover, while it may be that indivisibility results in socially valuable ideas failing to be produced, when new ideas are built on old ideas, government grants of intellectual monopoly may lead to even less innovation than under competition. The theory of the competitive provision of innovations we build is important both for understanding why in many current and historical markets there has been thriving innovation in the absence of copyrights, patents and other forms of monopoly power, and also for understanding why, in the presence of the rent-seeking behavior induced by government grants of monopoly, intellectual property in the form of copyrights and patents may be socially undesirable.

Classical economists believed the extent to which technological change may prevent the law of decreasing marginal productivity from taking its toll to be very limited. As economic growth continued at unprecedented rates, the central role of technological progress was recognized. With the notable exception of Schumpeter [1911], most early researchers either did not move past the narrative level or treated exogenous technological progress as a reasonable approximation.[1] Contributions by Lucas [1988] and Romer [1986] sparked a renewed attention to the theoretical issue. By developing and extending the arguments initially made by Arrow and Shell, these and other authors have argued that only models departing from the twin assumptions of decreasing returns to scale and perfect competition are capable of properly modeling persistent growth and endogenous technological progress. So, for example, Romer [1986] writes: "the key feature in the reversal of the standard results about growth is the assumption of increasing rather than

[1]There are few but important exceptions, which anticipated by a couple of decades subsequent developments. Most notably, Arrow [1962] and Shell [1966, 1967], to which we return later.

decreasing marginal productivity of the intangible capital good knowledge" (p. 1004).

Subsequent writings, such as Jones and Manuelli [1990] and Rebelo [1991], have pointed out that one can use a utility-maximizing version of the von Neumann [1937] model of constant returns to capture persistent growth. In such models, the linearity of the technology allows for unbounded accumulation of given capital goods. However, new commodities and new ways of producing them are not considered, either in theirs or subsequent works based on constant returns to scale technologies and unfettered competition. To study endogenous technological change,[2] most researchers have instead come to adopt models of monopolistic competition, such as the Dixit and Stiglitz [1977] model, and use increasing returns to describe the effect of technological change. It may not be an exaggeration to assert that a meaningful treatment of endogenous innovation and growth is commonly believed to be impossible under competitive conditions. Romer [1990a] asks, "Are Nonconvexities Important for Understanding Growth?" and answers with an unambiguous Yes.

We aim at disproving this belief. Our model can be interpreted as a positive theory of technological change in an economy in which legal monopoly rights are not conferred upon innovative entrepreneurs but in which there is a well defined "right of sale."[3] From an historical perspective, it seems unquestionable that the circumstances we model here have been the norm rather than the exception until, at least, the second half of the nineteenth century. Contemporary examples also abound and are illustrated below.

Endogenous economic innovation is the outcome of creative, purposeful effort. It is often argued that creative effort, the ideas it generates, and the goods in which it is embodied in must involve a fixed cost. Because of this, competitive markets are believed to be inconsistent with, or even harmful to, the development of new ideas. We cast doubt on such vision by arguing that a proper modeling of the production of ideas does not involve a fixed cost, but rather a sunk cost. There is little reason to believe that competition is unable to deal with sunk costs. The issue, if there is one, revolves around an indivisibility: half-baked ideas are seldom useful. Arrow [1962] points out the role of indivisibilities for understanding inventions (page 609), but his subsequent analysis concentrates mostly on inappropriability and uncertainty. Appropriability is addressed below. Uncertainty is ruled out by considering a deterministic environment, still it should be clear that the fundamental results can be replicated in a world where simple

[2]For the purpose of this paper, the expressions "technological change," "innovation," and "invention and adoption of new goods" should be taken as synonyms.

[3]A more precise definition of this concept is provided in Sections 2 and 3.

forms of uncertainty affected the innovation technology. Instead, we take on the study of indivisibilities from where Arrow left it: as a potential obstacle to competitive pricing of inventions. We conclude that this kind of indivisibility need not pose a substantive problem. This is akin to the observation made by Hellwig and Irmen [2001] that if the innovator has unique access to a strictly diminishing return technology and does not take advantage of his monopoly over production, never-the-less innovation will occur. However, Hellwig and Irmen maintain both a production technology that involves a fixed cost and, more importantly, the assumption that ideas, after some delay, "spillover" without cost or without the necessity of paying for either the idea or for a person or good in which it is embodied. Because, as we argue below, ideas are embodied and costly to transmit, we do not think spillovers are an important externality. Because we do not allow spillover, unlike Hellwig and Irmen [2001], we can identify circumstances under which competitive equilibrium yields the first best outcome.

There is an influential literature, advocating a close connection between innovative activity and the establishment of monopoly rights (Aghion and Howitt [1992], Grossman and Helpman [1991], Romer [1990a,b]). In this setting, new goods and new technologies are introduced because of the role of individual entrepreneurs in seeking out profitable opportunities. Such profitable opportunities arise from monopoly power. We too consider the role of entrepreneurship in seeking out profitable opportunities, but unlike this early literature, we do not assume monopolistic competition or increasing returns to scale. When there is no indivisibility, our technology set is a convex cone and competitive equilibria are efficient. Technological progress takes place because entrepreneurs find it advantageous to discover and produce new commodities. These new commodities themselves may make profitable the employment of new activities that make use of them. Although, in the ensuing equilibrium, entrepreneurs do not actually end up with a profit, it is their pursuit of profit that drives innovation.

The central feature of any story of innovation is that rents, arising from marginal values, do not fully reflect total social surplus. This may be due to non rivalry or to an indivisibility or to a lack of full appropriability. Non rivalry we discuss thoroughly in the next section. Appropriability, or lack of it thereof, depends on whether ideas can be obtained without paying the current owner. Romer [1990a] argues convincingly that appropriability (excludability in his terminology) has no bearing on the shape of the feasible technology set. Since we do not believe that ideas are easily obtained without paying at least for goods that embody them, we do not believe that appropriability is an important problem. In our analysis we assume full appropriability of privately produced commodities and concentrate on the presence of an indivisibility in the inventive process. With indivisibility in

production total surplus matters, not rents, so competitive economies may fail to produce socially desirable innovations. We do not disagree with this assessment. We do wish to shed doubt on how important it is, both in principle and in reality, and on whether government enforced monopoly is a sensible response to the problems it involves. First, in many practical instances, rents are adequate to pay for the cost of innovation, and lowering reproduction costs does not generally reduce, indeed: it often increases, such incentive. Second, while awarding a monopoly to an innovator increases the payoff to the original innovator, by giving her control over subsequent uses of the innovation, it reduces the incentive for future innovation. This point has been strongly emphasized by Scotchmer [1991]. In our setting, we show how indivisibility may lead monopoly to innovate less than competition. Hence, we argue, our analysis has normative implications for those markets in which innovative activity satisfies the assumptions of the model presented here. As a further application of our positive theory, we consider the impact of more efficient technologies for the reproduction of ideas on the large rents that may accrue to superstars, even in the absence of monopoly.

Historically, then, we believe that the theory of innovation under competition is important for understanding growth and development, since government intervention in the market for ideas is a relatively recent development. Since we establish that there are economies in which competition without patents and copyright achieves the first best, the issue of whether government grants of monopolies over ideas is second best is an empirical rather than theoretical issue. There are few empirical studies that shed light on this question. There is a great deal of less formal evidence that shows that innovation can thrive under competition; and that government grants of monopoly power are more prone to lead to socially costly rent-seeking behavior than to foster innovation and growth. Since we feel that there should be a presumption against government grants of monopoly, we give three examples. The first concerns copyright. From Arnold Plant [1934] we learn that "During the nineteenth century anyone was free in the United States to reprint a foreign publication" without making any payment to the author. This is a fact that greatly upset Charles Dickens whose works, along with those of many other English authors, were widely distributed in the U.S. And "yet American publishers found it profitable to make arrangements with English authors. Evidence before the 1876-8 Commission shows that English authors sometimes received more from the sale of their books by American publishers, where they had no copyright, than from their royalties in [England]." The second concerns patentable ideas. From George Stigler [1956] we learn "There can be rewards - and great ones - to the successful competitive innovator. For example, the mail-order business was

an innovation that had a vast effect upon retailing in rural and small urban communities in the United States. The innovators, I suppose, were Aaron Montgomery Ward, who opened the first general merchandise establishment in 1872, and Richard Sears, who entered the industry fourteen years later. Sears soon lifted his company to a dominant position by his magnificent merchandising talents, and he obtained a modest fortune, and his partner Rosenwald an immodest one. At no time were there any conventional monopolistic practices [or patents], and at all times there were rivals within the industry and other industries making near-perfect substitutes (e.g. department stores, local merchants), so the price fixing-power of the large companies was very small." One of the few studies that argues that there is evidence that patents do promote innovation is a study of patents in the late 19th and early 20th century by Lamoreaux and Sokoloff [2002]. They argue that innovation expanded greatly as a consequence of a change in the patent system that took place in 1836. What was the change that led to this explosion? Under the 1836 legislation "technical experts scrutinized applications for novelty and for the appropriateness of claims about invention." In other words, the change that led to the explosion of innovation was a legal change that made it *more* difficult to get a patent. This observation also contains a cautionary note for those who believe that a tightly run patent system is good for innovation - patent offices are as prone to regulatory capture as any other government agency. Today the patent office will apparently patent anything regardless of novelty and merit: the list of silly patents in recent years includes among other things, a patent on swinging on a swing; the peanut butter and jelly sandwich, and a method of transmitting energy by poking a hole in another dimension. Moreover, patents have recently expanded to include "business practices" including financial securities. This despite Tofuno's [1989] careful documentation of the enormous amount of innovation that took place in the financial industry under a competitive environment such as the one we study in this paper.

2. PRICING OF IDEAS

It is widely accepted that every process of economic innovation is characterized by two phases. First comes the research and development or invention step, aimed at developing the new good or process; second comes the stage of mass production, in which many copies of the initial prototype are reproduced and distributed. The first stage is subject to a minimum size requirement: given a target quality for the new product or process, at least one prototype must be manufactured. Such a minimum size requirement corresponds to an initial indivisibility: there exists a strictly positive lower bound on the amount of resources to be devoted to any inventive process.

After the invention stage is completed and some goods embodying the new idea are produced, large scale replication takes place at a low and practically constant marginal cost. To avoid future misunderstandings, let us stress here that the expression "goods embodying the new idea" may mean either the new good in the strictest sense, or a set of capital tools needed to produce it, or a body of knowledge embodied in people and goods needed to replicate the innovation.

We agree with this popular description. The contrast between the invention and reproduction stages can be made sharper by pointing at the extreme case in which, after the invention is completed, it is the new idea itself that is being reproduced and distributed. Indeed, in the case of artistic works, for example, it is only the production and distribution of the message (idea) that matters, not the media through which it goes. (The medium is not the message.) One model of the production and distribution of ideas is to assume that they take place with an initial fixed cost. The technical description is that ideas are nonrivalrous: once they exist they can be freely appropriated by other entrepreneurs. Since at least Shell [1966, 1967], this is the fundamental assumption underlying the increasing returns-monopolistic competition approach: "technical knowledge can be used by many economic units without altering its character" (Shell [1967, p. 68]). Our use of the fundamental theorem of calculus cannot prevent innumerable other people from using the same theorem at the same time. While this observation is correct, we depart from conventional wisdom because we believe it is irrelevant for the economics of innovation. What is economically relevant is not some bodyless object called *the fundamental theorem of calculus*, but rather our personal knowledge of the fundamental theorem of calculus. Only ideas embodied in people, machines or goods have economic value. To put it differently: economic innovation is almost never about the adoption of new ideas. It is about the production of goods and processes embodying new ideas. Ideas that are not embodied in some good or person are not relevant. This is obvious for all those marvelous ideas we have not yet discovered or we have discovered and forgotten: lacking embodiment either in goods or people they have no economic existence. Careful inspection shows the same is also true for ideas already discovered and currently in use: they have economic value only to the extent that they are embodied into either something or someone. Our model explores the implications of this simple observation leading to a rejection of the long established wisdom, according to which "for the economy in which technical knowledge is a commodity, the basic premises of classical welfare economics are violated, and the optimality of the competitive mechanism is not assured." (Shell [1967, p.68]). In short, we reject the idea of unpriced "spillovers." Regardless of the legal framework, no inventor of an idea is obligated to share his idea with others

for free. It may, of course, be considerably more expensive to come up with a good new idea than it is to buy a product embodying the idea and copy it. This, however, is not an unpriced "spillover" and need not necessarily pose a disincentive to coming up with new ideas.

A couple of additional examples may help clarify the intuition behind our modeling strategy. Take the classical and abused case of a software program. To write and test the first version of the code requires a large investment of time and resources. This is the cost of invention mentioned before, which is sunk once the first prototype has been produced. The prototype, though, does not sit on thin air. To be used by other it needs to be copied, which requires resources of various kinds, including time. To be usable it needs to reside on some portion of the memory of your computer. To put it there also requires time and resources. If other people want to use the original code to develop new software, they need to acquire a copy and then either learn or reverse-engineer the code. Once again, there is no free lunch: valuable ideas are embodied in either goods or people, and they are as rivalrous as commodities containing no ideas at all, if such exist. In our view, these observations cast doubts upon Romer's [1986, 1990a, 1990b] influential argument according to which the nonrivalrous nature of ideas and their positive role in production *a fortiori* imply that the aggregate production function displays increasing returns to scale. A stylized representation of these different views about the production function for idea goods is in Figure 1. In one case, shown by the thick line, there is a fixed cost: input levels less than or equal to $\underline{h} > 0$ yield zero output. From $\underline{h}$, the technology is one of constant returns; as a consequence the aggregate technology set is not convex. This is the established view. In the alternative case, shown by the thin line, there is an indivisibility: if strictly less than $\underline{h}$ units of input are invested, there is no output. When the critical level $\underline{h}$ is reached the first (or first few) units of the new good are produced. After that the common hypothesis of constant returns to scale holds. In the latter case the aggregate technology set is convex when the minimum size requirement is not binding. This is the theory being proposed here, Our contention is that the

latter is a *more appropriate* representation of the innovation process than

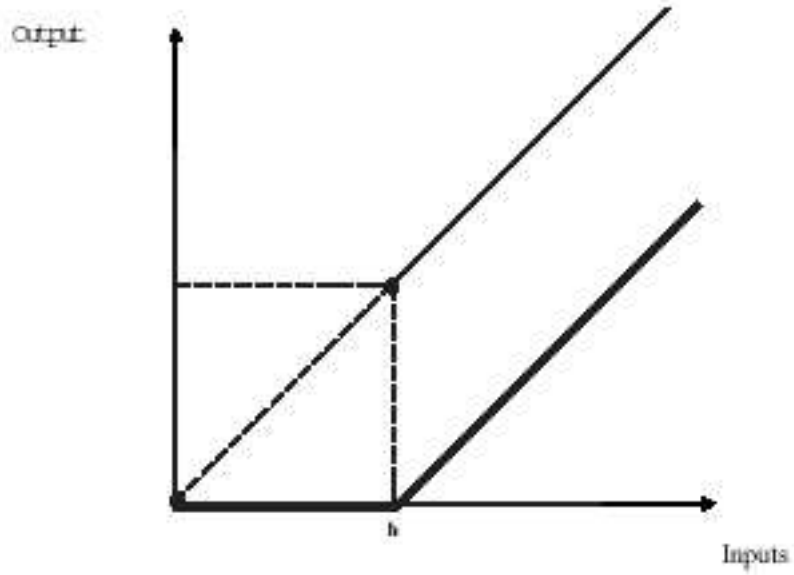

Figure 1: Fixed costs vs. indivisibility

the former.

Proponents of the standard view observe that "Typically, technical knowledge is very durable and the cost of transmission is small in comparison to the cost of production" (Shell [1967, p. 68]). Admittedly, there exist circumstances in which the degree of rivalry is small, almost infinitesimal. Consider the paradigmatic example of the wheel. Once the first wheel was produced, imitation could take place at a cost orders of magnitude smaller. But even imitation cannot generate free goods: to make a new wheel, one needs to spend some time looking at the first one and learning how to carve it. This makes the first wheel a lot more valuable than the second, and the second more valuable than the hundredth. Which is a fine observation coinciding, verbatim, with a key prediction of our model. The "large cost of invention and small cost of replication" argument does not imply that the wheel, first or last that it be, is a nonrivalrous good. It only implies that, for some goods, replication costs are very small. If replication costs are truly so small, would it not be a reasonable approximation to set them equal to zero and work under the assumption that ideas are nonrivalrous? Maybe. As a rule of scientific endeavor, we find approximations acceptable when their predictions are unaffected by small perturbations. Hence,

conventional wisdom would be supported if perturbing the nonrivalry hypothesis did not make a difference with the final result. As we show, it does: even a minuscule amount of rivalry can turn standard results upside down.

3. INNOVATION UNDER COMPETITION

The list of all goods that conceivably can be produced is a datum. So are the procedures (activities, in our language) through which goods can be obtained. Very many conceivable goods and activities, indeed most of them, are not produced or used at any point in time. For the purpose of this article, an (economic) *innovation* is therefore defined as the first time a good is actually produced or an activity is employed.

To understand whether an innovation will take place or not in a competitive environment, we must understand how much a new good is worth after it is created. Consider a competitive market in which an innovation has already been produced. In other words, there is currently a single template item, book, song, or blueprint that is owned by the creator. We focus on the extreme case where every subsequent item produced using the template is a perfect substitute for the template itself - that is, what is socially valuable about the invention is entirely embodied in the product.[4] At a moment in time, each item has two alternative uses: it may be consumed or it may be used to produce additional copies. For simplicity we assume that while the process of copying is time consuming, there is no other cost of producing copies.

Specifically, suppose that there are currently $k > 0$ units of the innovative product available. Suppose that $0 < c \leq k$ units are allocated to consumption, leaving $k - c$ units available for the production of copies. The $k - c$ units that are copied result in $\beta(k - c)$ copies available in the following period, where $\beta > 1$. Because the units of the good used in consumption might be durable, there are ζc additional units available next period. In many cases $\zeta \leq 1$ due to depreciation, however we allow the possibility that the good may be reproduced while consumed, and require only that it not be easier to reproduce while consuming $\zeta \leq \beta$. The case in which reproduction does not interfere with consumption, that is, the Quah [2002] 24/7 case in which $\zeta = \beta$ will be explicitly considered below.

The representative consumer receives a utility of $u(c)$ from consumption, where u is strictly increasing, concave, and bounded below. The infinitely lived representative consumer discounts the future with the discount factor

[4]Notice that the "product" could be a book, or a progress report, or an engineering drawing of a new production process containing detailed instructions for its implementation, or just a collection of people that have learned how to "do it".

$0 \leq \delta < 1$. We assume that the technology and preferences are such that feasible utility is bounded above.

It is well known that the solution to this optimization problem may be characterized by a concave value function $v(k)$, which is the unique solution of

$$v(k) = \max_{0 \leq c \leq k} \{u(c) + \delta v(\beta k - (\beta - \zeta)c)\}.$$

In an infinite horizon setting, beginning with the initial stock of the new good $k_0 = k$ we may use this program recursively to compute the optimal k_t for all subsequent t. Moreover, the solution of this problem may be decentralized as a competitive equilibrium, in which the price of consumption services in period t is given by $p_t = u'(c_t)$. >From the resource constraint

$$c_t = \frac{\beta k_t - k_{t+1}}{\beta - \zeta}.$$

If ζ is large enough relative to β it may be optimal not to invest at all and to reproduce solely by consuming. We first take the case where consumption is strictly less than capital in every period. By standard dynamic programming arguments, the price q_t of the durable good k_t can be computed as

$$q_t = v'(k_t) = p_t \frac{\beta}{\beta - \zeta}.$$

As $p_t > 0$, $q_t > 0$ for all t. The zero profit condition implies that q_t decreases at a rate of $1/\beta$ per period of time.

Consider then the problem of innovation. After the innovation has occurred, the innovator has a single unit of the new product $k_0 = 1$ that he must sell into a competitive market: there is no patent or copyright protection. In a competitive market the initial unit sells for q_0, which may be interpreted as the rent accruing to the fixed factor $k_0 = 1$ owned by the innovative entrepreneur. The market value of the innovation corresponds, therefore, to the market value of the first unit of the new product. This equals, in turn, the net discounted value of the future stream of consumption services it generates. This is what we mean with the right of sale, from which rents accruing to competitive innovators originate. Introducing that first unit of the new good entails some cost $C > 0$ for the innovator. Consequently, the innovation will be produced if and only if the cost of creating the innovation C is less than or equal to the rent resulting from the innovation and captured by the fixed factor, $C \leq q_0$..

Notice that $q_0 \geq p_0 = u'(c_0) \geq u'(1)$. The first inequality is strict whenever $\zeta > 0$. Notice also that there is no upper bound on the number of units of the new good that can be produced and that there is no additional cost of making copies. Indeed, the only difference between this model and the

model in which innovations are nonrivalrous is that in this model, as in reality, reproduction is time consuming and, as in reality again, there is an upper bound $\beta < \infty$ on how many copies may be produced per unit of time, that is, there is limited capacity at any given point in time. These twin assumptions capture the observation, discussed earlier, that nonrivalry is only an approximation to the fact that costs of reproduction are very small. Consequently, this simple analysis clarifies that there can be no question that innovation can occur under conditions of perfect competition.

A less obvious question is: What happens as β, the rate at which copies can be made, increases? If, for example, the advent of the Internet makes it possible to put vastly more copies than in the past in the hands of consumers in any given time interval, what would happen to innovations in the absence of legal monopoly protection? Conventional wisdom suggests that in this case, rents fall to zero, and competition must necessarily fail to produce innovations. This conclusion is basically founded upon examination of a static model with fixed cost of invention and no cost of reproduction. As we shall see, conventional wisdom fails for two reasons: first, it ignores the initial period. During this initial period, no matter how good the reproduction technology, only one copy is available. In other words, $q_0 \geq u'(c_0)$ is bounded below by $u'(1)$ regardless of β and of the speed of depreciation. With impatient consumers, the amount that will be paid for a portion of the initial copy (or, more realistically, for one of the few initial specimens of the new good) will never fall to zero, no matter how many copies will be available in the immediate future. This consideration has great practical relevance for markets such as those for artistic works, where the opportunity to appreciate the work earlier rather than later has great value. The very same argument applies, even more strongly, to innovative medical treatments. Empirical evidence suggests that getting there earlier has substantial value in other highly innovative industries, such as the financial securities industry (Schroth and Herrera [2001]). In other words, regardless of copyright law, movies will continue to be produced as long as first run theatrical profits are sufficient to cover production costs; music will continue to be produced as long as profits from live performances are sufficient to cover production costs, books will continue to be produced as long as initial hardcover sales are sufficient to cover production costs, and financial and medical innovations will take place as long as the additional rents accruing to the first comer compensate for the R&D costs.

Conventional wisdom also fails for a second, less apparent, reason: increasing β may increase, rather than decrease, the rent to the fixed factor.

Observe that $\partial q_0/\partial p_0 > 0$ and that

$$\frac{dq_0}{d\beta} = u''(c_0)\frac{dc_0}{d\beta} - u'(c_0)\frac{\zeta}{(\beta-\zeta)^2}.$$

When β is sufficiently large relative to ζ the first term will dominate. For concreteness, consider the case of full depreciation, $\zeta = 0$. In this case the rent will increase if initial-period consumption falls with β and will decrease if it rises. In other words, the relevant question is whether consumptions are substitutes or complements between time periods. If they are substitutes, then increasing β lowers the cost of consuming after the first period and causes first period consumption to decline to take advantage of the reduced cost of copies in subsequent periods. This will increase the rent to the fixed factor and improve the chances that the innovation will take place. Conversely, if there is complementarity in consumption between periods, the reduced cost in subsequent periods will increase first-period consumption of the product and lower the rent.

It is instructive to consider the case in which the utility function has the CES form $u(c) = -(1/\theta)(c)^{-\theta}$, $\theta > -1$.[5] In this case, it is possible to explicitly compute the optimal consumption/production plan. Consider first the case of inelastic demand where $\theta > 0$. Here there is little substitutability between periods and a calculation shows that as $\beta \to \infty$ initial consumption $c_0 \to \overline{c} < 1$. Consequently, rents from innovation fall, but not toward zero. Competitive innovation still takes place if $\overline{p} = u'(\overline{c}) > C$.

More interesting is the case of elastic demand, where $\theta \in (-1, 0]$. This implies a high elasticity of intertemporal substitution in consumption ($\theta = -1$ corresponds to linear utility and perfect substitutability). Utility becomes unbounded above as $\beta \to \delta^{1/\theta}$. A simple calculation shows that as this limit is approached, $c_0 \to 0$ and rents to innovators become infinite. In other words, in the CES case, with elastic demand, every socially desirable innovation will occur if the cost of reproduction is sufficiently small. This case is especially significant, because it runs so strongly against conventional wisdom: as the rate of reproduction increases, the competitive rents increase, despite the fact that over time many more copies of the new good are reproduced and distributed. This is true even if, following Quah's 24/7 model, we assume that the initial time period is arbitrarily short. So the basic assumptions is simply that demand for the new product is elastic. Notice that currently accepted theories argue, as do current holders of monopoly rights, that, with the advent of a technology for cheap reproduction, innovators' profits are threatened and increased legal monopoly powers are

[5] Strictly speaking, we assume CES utility above a certain minimum subsistence level of consumption.

required to keep technological innovation from faltering. This model shows that quite the opposite is possible: decreasing the reproduction cost makes it easier, not harder, for a competitive industry to recover production costs.

So far we have assumed that depreciation is sufficient to give an interior solution. It is instructive also to consider the polar opposite case in which $\zeta = \beta$ and reproduction does not interfere with consumption. In this case, capital grows at the fixed exogenous rate of β and the initial rent is given by

$$q_t = \sum_{i=0}^{\infty} (\beta\delta)^i p_{t+i} = \sum_{i=0}^{\infty} (\beta\delta)^i u'(\beta^i k_t).$$

In particular, the basic feature thats rents are bounded below by the marginal utility of the initial unit, and that in the elastic case the initial rent becomes infinite as the reproduction rate increases remain true.

4. Innovation Chains

A central feature of innovation and growth is that innovations generally build on existing goods, that is, on earlier innovations. Scotchmer (1991) has particularly emphasized this feature of innovation. We now extend the theory of the previous section to consider a situation where each innovation creates the possibility of further innovations. We focus first on a positive theory of the role of indivisibility in competitive equilibrium. In contrast to the previous section, we now assume that there are many producible qualities of capital, beginning with quality zero. We denote capital of quality i by k_i. As before, capital may be allocated to either consumption or investment. Each unit of capital of quality i allocated to the production of consumption yields $(\gamma)^i$ units of output where $\gamma > 1$, reflecting the greater efficiency of higher quality capital. As before, capital used to produce consumption is assumed to depreciate at the rate $1-\zeta$. Suppose that c_i units of consumption are produced from quality i capital. This leaves $k_i - c_i/\gamma^i$ units available for investment. As before each unit of capital may be used to produce $\beta > 1$ copies of itself. However, we now assume that capital may also be used as an input into the production of higher quality capital. Specifically, if h_i units of capital are allocated to innovation, ρh_i units of quality $i+1$ capital result next period. Because innovation is costly, we assume $\rho < \beta$. Because half-baked new goods are of no use, an indivisibility may characterize the process of innovation, so that a minimum of $\underline{h} \geq 0$ units of capital must be invested before any output is achieved using the ρ technology. In the context of innovation chains, the indivisibility plays the role that the large cost of innovation C played in the one-shot innovation model of the previous section. Repeated innovation takes place only if rents from the introduction

of capital $i+1$ are large enough to compensate for investing at least $\underline{h}$ units of capital i in the innovation process.

The only interesting case is the one in which $\rho\gamma > \beta$ so that innovation is socially desirable. Moreover, as our focus is on growth rather than decline, we will assume that technology is productive enough to yield sustained growth. Observe that, independent of depreciation, giving up a unit of consumption today always yields a net gain of β units of consumption tomorrow. We assume

Assumption 4.1. $\delta\beta > 1$.

This assumption means that by using the β technology it is possible for the capital stock to grow faster than the inverse discount factor.

4.1. **Convex Production Possibilities.** To analyze competition in this setting, it is useful to begin by considering the standard case of a convex production set, in which $\underline{h} = 0$.

When $\rho\gamma > \beta$, the technology of producing copies using the β activity is dominated by the technology of innovating using the ρ activity. This implies that the β activity is never used. However, at any moment of time, there will typically be several qualities of capital available: the new qualities produced through the innovation and the old qualities left over after depreciation. It is important to note that, in the absence of the minimum size restriction, if several qualities of capital are available at a moment in time, it is irrelevant which quality is used to produce consumption: the trade-off between consumption today and tomorrow is the same for all qualities of capital. Hence, the quality composition of capital does not affect the rate of technology adoption and consumption growth in the absence of indivisibilities. In particular, let "consumption productive capacity" in period t be defined as[6]

$$P_t = \sum_{i=0}^{I_t} \gamma^i k_t^i$$

where I_t is the set of all kinds of capital available in period t. Observe that, no matter which among the components of the vector $\left[k_t^0, k_t^1, ..., k_t^{I_t}\right]$ is used to produce consumption in period c_t and which is used to produce P_{t+1}, the tradeoff between c_t and P_{t+1} is always at the rate $\rho\gamma$.. Hence, in competitive equilibrium, consumption must satisfy the first-order condition that the marginal rate of substitution equals the marginal rate of transformation:

$$u'(c_t) = \delta\rho\gamma u'(c_{t+1}).$$

[6]Our most sincere thanks to Jim Dolmas for patiently forcing our stubborn minds to see an algebraic mistake in an earlier derivation of these conditions.

Since u is strictly concave, u' is strictly decreasing, and this immediately implies that $c_{t+1} > c_t$, that is, there is continued growth if and only if $\delta\rho\gamma > 1$. Suppose also that we make a modest regularity assumption on preferences

Assumption 4.2. The coefficient of relative risk aversion $-cu''(c)/u'(c)$ is bounded above and bounded away from zero as $c \to \infty$.

Notice that this is true for all utility functions that exhibit nonincreasing absolute risk aversion and, in particular, for all CES utility functions. Under this assumption, we may conclude from applying Taylor's theorem to the first-order condition above that not only is $c_{t+1} > c_t$, but, in fact, $(c_{t+1} - c_t)/c_t > \Delta > 0$ and, in particular, c_t grows without bound. Hence, repeated competitive innovations take place because rents are high enough to provide an incentive for entrepreneurs to undertake the innovative activity.

4.2. **Growth with Indivisibility.** We now consider the case with an indivisibility $\underline{h} > 0$. Clearly, if the indivisibility is large enough, competitive equilibrium in the usual sense may not exist. However, if the indivisibility is not so large, it may not bind at the social optimum, in which case the usual welfare theorems continue to hold, and the competitive equilibrium provides a continuing chain of innovations.

In considering the role of indivisibilities in the innovation process, a key question is, What happens to investment in the newest technology over time, that is, to the amount of resources allocated to technological innovation? If it declines to zero, then regardless of how small $\underline{h}$ is, the indivisibility must eventually bind. Conversely, if the investment grows or remains constant, then a sufficiently small $\underline{h}$ will not bind. Notice that for any particular time horizon, since consumption is growing over time, investment is always positive, so a small enough $\underline{h}$ will not bind over that horizon. Consequently, we examine what happens asymptotically to investment in the newest quality of capital.

We study asymptotic investment by making the assumption that for large enough c the utility function $u(c)$ has approximately the CES form $u(c) = -(1/\theta)(c)^{-\theta}$, $\theta > -1$. In the CES case, we can explicitly solve the first-order condition from above to find the growth rate of consumption g as

$$g = \frac{c_{t+1}}{c_t} = (\delta\rho\gamma)^{1/(1+\theta)}.$$

Notice that without the indivisibility, it makes no difference whether old and depreciated or newly produced capital is used to produce consumption. In other words, as already noted earlier, the quality composition of capital does not matter for the equilibrium path in the absence of the indivisibility. This is no longer true with the indivisibility, since it may be that there

are many different production plans that, by using different combinations of capital of different qualities, achieve the growth rate of consumption given above. Notice, for example, that when capital of quality i is introduced from capital of quality $i-1$, the amount available after the first round may not be enough to immediately exceed the threshold $\underline{h}$ needed for the introduction of quality $i+1$ capital. Still, there may be enough newly produced capital to meet the consumption target in that period while, at the same time, there is sufficient depreciated old capital of type $i-1$ to produce additional capital of type i to pass the innovation threshold next period. In this specific example, then, consumption grows at the rate g defined by the unconstrained first-order conditions, while a new quality of capital is introduced only every second period. Things are even more complex in those cases in which the optimal plan calls for using the ρ technology in certain periods to introduce new qualities of capital and the β technology in other periods to accumulate capital faster until the threshold level $\underline{h}$ is reached. While alternating periods of capital widening and capital deepening may be a fascinating theoretical scenario to investigate, because they resemble so much what we observe in reality, these complications make a full characterization of the equilibrium production plan beyond the scope of the present paper.[7]

For the present analysis it will suffice to notice that, if our growth condition is satisfied, it is likely to be satisfied strictly, meaning that investment in the newest quality of capital grows asymptotically exponentially when measured in physical units. This implies that the indivisibility is binding only earlier on and becomes irrelevant after a finite number of periods, as the threshold $\underline{h}$ is vastly exceeded.[8] In other words, as the scale of physical capital increases, the quantity devoted to innovation increases, and the problem of minimal scale becomes irrelevant. Put in terms of innovation, this says that as the stock of capital increases, rather than a single innovation, we should expect many simultaneous innovations in any given period. In fact, cases of simultaneous discovery seem to be increasingly frequent in advanced economies as the amount of resources devoted to R&D increases.

[7] A model of endogenous growth through oscillations between innovation and accumulation is in Boldrin and Levine [2002a].

[8] Indivisibilites may therefore play a crucial role in the early stages of economic growth. This is the central theme of Acemoglu and Zilibotti [1997] who, in a context of uncertain returns from different investment projects, concentrate on the role of indivisibilities as the key obstacle to optimal diversification. They show that indivisibilities may retard economic growth in poor countries. Our analysis upports their findings as it suggests that indivisibility also hampers innovations and technological progress. Notice, in passing, that both in the Acemoglu and Zilibotti's model and in ours free mobility of capital facilitates economic growth; in our model free trade of final commodities, by itself, may facilitate technological change and economic growth (Boldrin and Levine [2003].)

It can be argued that this is in part due to patent law, which rewards first past the post, inducing patent races. However, it should be noted that rapid and parallel development occur frequently without the benefit of patent protection. This is the case of basic science, where patent law is not applicable, and also in the case of open source software development, where the innovators choose not to protect their intellectual property through restrictive downstream licensing agreements. The fashion industry, where labels are protected but actual designs are frequently replicated at relatively low costs (e.g. the *Zara* phenomenon) is another striking example.

4.3. **Entrepreneurship, Profits, and Competition.** In our competitive setting, entrepreneurs have well-defined property rights to their innovations, individual production processes display constant returns, and there are no fixed costs and no unpriced spillover effects from innovation. Entrepreneurs also have no ability to introduce monopoly distortions into pricing. Does this lead to an interesting theory of innovation? We believe it provides a positive theory of the many thriving markets in which innovation takes place under competitive conditions. In addition to the examples of fashion, open software, and basic scientific knowledge already mentioned, there are a variety of other thriving markets that are both competitive and innovative, such as the market for pornography, for news, for advertising, for architectural and civil engineering designs, and, for the moment at least, for recorded music. A particularly startling example is the market for financial securities. This was documented by Tofuno [1989], and, more recently, by Schroth and Herrera [2001]. Their empirical findings document that despite the absence of patent and copyright protection and the extremely rapid copying of new securities, the original innovators maintain a dominant market share by means of the greater expertise they have obtained through innovation. Maybe less scientifically compelling, but not less convincing, is the evidence reported by Lewis [1989] and Varnedoe [1990].[9] They provide vivid documentation of the patterns of inventive activity in, respectively, investment banking and modern figurative arts, two very competitive sectors in which legally enforced monopoly of ideas is altogether absent.

Although the basic ingredients of our theory of fixed factors, rents, and sunk costs are already familiar from the standard model of competitive equilibrium at least since the work of Marshall, the way in which they fit together in an environment of growth and innovation is apparently not well understood. Central to our analysis is the idea that a single entrepreneur contemplating an innovation anticipates the prices at which he will be able to buy inputs and sell his output and introduces the innovation if, at those

[9]We owe the first suggestion to Pierre Andre Chiappori and the second to Robert Becker.

prices, he can command a premium over alternative uses of his endowment. He owns the rights to his innovation, meaning that he expects to be able to collect the present discounted value of downstream marginal benefits. As we have shown, this provides abundant incentives for competitive innovation.

In the model of innovation chains, an entrepreneur who attempted to reproduce his existing capital of quality i when the same capital can be used to introduce capital of quality $i+1$ would make a negative profit at equilibrium prices. In this sense, the competitive pressure from other entrepreneurs forces each one to innovate in order to avoid a loss.

As in theories of monopolistic competition and other theories of innovation, new technologies are introduced because of the role of individual entrepreneurs in seeking out profitable opportunities. Unlike in those theories, the entrepreneur does not actually end up with a profit. Because of competition, only the owners of factors that are in fixed supply can earn a rent in equilibrium. When a valuable innovation is introduced, it will use some factors that are in fixed supply in that period. Those factors will earn rents. If you are good at writing operating systems code when the personal computer technology is introduced, you may end up earning huge rents, indeed. In principle, this model allows a separation between the entrepreneurs who drive technological change by introducing new activities and the owners of fixed factors who profit from their introduction. However, it is likely in practice that they are the same people.

5. Does Monopoly Innovate More than Competition?

Conventional economic wisdom argues that innovation involves a fixed cost for the production of a nonrivalrous good. That is to say, there are increasing returns to scale due to the role of ideas in the aggregate production function. It is widely believed that competition cannot thrive in the face of increasing returns to scale, and so the discussion quickly moves on to other topics: monopolistic competition, government subsidy, or government grants of monopoly power. We have argued in the previous sections that this conventional wisdom is misguided. Innovation involves a sunk cost, not a fixed cost, and because ideas are embodied in people or things, all economically useful production is rivalrous. Sunk costs, unlike fixed costs, pose no particular problem for competition; indeed, it is only the indivisibility involved in the creation of new ideas that can potentially thwart the allocational efficiency of competitive prices. In the end, it is necessary only that the rent accruing to the fixed factors comprising the new idea or creation cover the initial production cost. When innovations feed on previous ones, we have shown that in many cases the increasing scale

of investment in R&D leads over time to many simultaneous ideas and creations, thereby making the indivisibility irrelevant. In short, we have argued that the competitive mechanism can be a viable one, capable of producing sustained innovation.

This is not to argue that competition is the best mechanism in all circumstances. In fact, rents to a fixed factor may fall short of the cost of producing it, even when the total social surplus is positive. Indivisibility constraints may bind, invalidating the analysis of the previous sections. Nevertheless, even in this case we do not find it legitimate to conclude that competition fails. More appropriately, we simply gather from this that we do not yet have an adequate theory of competitive equilibrium when indivisibility constraints bind. Could, for example, clever entrepreneurs eke out enough profit in a competitive environment in which traditional rents do not cover innovation costs by taking contingent orders in advance, or by selling tickets to a lottery involving innovation as one outcome? Entrepreneurs have adopted exactly such methods for many centuries in markets where indivisibilities have posed a problem. In the medieval period, the need for convoys created a substantial indivisibility for merchants that was overcome through the clever use of contingent contracts. In modern times, Asian immigrants (among other) have overcome the need for a minimum investment to start a small business by organizing small lottery clubs.

We do not have a positive theory of competitive markets when the indivisibility constraint binds and innovation is recursive. Can there be a competitive equilibrium in which innovation is delayed in order to accumulate enough capital to overcome the indivisibility? What are the welfare consequences of competitive equilibrium? We do not know the answer to these questions. What we do know is that competition is a powerful force and that entrepreneurs are generally more creative than economic theorists. Few advocates of monopoly rights, we suspect, would have predicted that a thriving industry of radio and television could be founded on the basis of giving the product away for free.

Let us accept, however, that under the competitive mechanism, some socially desirable innovations and creations will not be produced. Can this be overcome by government grants of monopoly to producers of innovations and creations? Conventional wisdom says that a monopolist can recover no less profit than competitors, and so is at least as likely to cover innovation costs. This picture of the monopolist as aggressive innovator may come as a shock to noneconomists and empiricists, but underlies the literature on patent and copyright protection. The problem is this: while giving monopoly rights to an innovator enhances his incentive to innovate at a given point in time, it is also likely to create incentives to suppress all subsequent innovations. Consequently, grants of monopoly rights not only create monopoly

distortions for innovations that would have taken place anyway, but may lead to less, rather than more, innovation. This danger of monopoly when innovations build on past innovations has been emphasized by Scotchmer [1991]. The very same danger exists in our setting as well.

To model dynamic monopoly in the setting of innovation chains poses a number of complications. Because issues of commitment, timing, and the number of players matter in a game played between a long-run monopolist and non-atomistic consumers or innovators, we must take greater care in specifying the environment than in the case of competition. We are not aiming here at a general theory of monopolistic behavior in the presence of innovation chains. Our goal is simply to expose the retardant effect that legally supported monopoly power may have upon the rate of technological innovation. Specifically, we make the following assumptions. Retain the set of commodities and activities from the previous sections, and add a transferable commodity m. Assume next a transferable utility model, meaning that consumer utility is $m + \sum_{t=0}^{\infty} \delta^t u(c_t)$ and that the utility of the monopolist is simply m. Initially the consumer is endowed with a large amount $\overline{m}$ of the transferable commodity, while the monopolist is endowed with none. In addition, we assume that at the beginning of each period, the monopolist chooses a particular production plan and that the price for consumption is subsequently determined by consumers' willingness to pay. Finally, we assume that the monopolist owns the initial capital stock (k_0^0) and has a complete monopoly over every output produced directly or indirectly from his initial holding of capital. In other words, beside owning the stock of capital the monopolist has also been awarded full patent protection over the β, ρ and γ activities that use that capital as an input. This leads to a "traditional" model of monopoly in the sense that consumers are completely passive, and there is a unique equilibrium in which precommitment makes no difference.

Of these assumptions, we should single out the assumption that the monopolist controls all production, either direct or indirect, from his original innovation. In particular, we assume that the monopolist not only can prevent consumers from employing the β technology to reproduce copies of the work, but can also prevent them from using the ρ technology to produce innovations of their own. We should note that this is a more extreme form of monopoly than that envisaged under current U.S. law on intellectual property. Patent law, on the one hand, gives the innovator complete control over the uses of the innovation, but only for 20 years, and there may be practical problems in showing that a particular patented idea was used in the production of another idea. Copyright, by way of contrast gives rights that

effectively last forever,[10] but until the passage of the Digital Millennium Copyright Act in 1998, allowed the consumer the right of "fair use." At the current time, for example, a copyright holder has rights over sequels to her works, but not over parodies. As in the case of patent law, it may difficult in practice to enforce these rights.

Our goal is a fairly specific one: to show that a monopolist who has complete downstream rights may have an incentive to suppress innovation, even in circumstances where a competitive industry would innovate. In particular, we construct an example in which we begin with a situation where there is no indivisibility, so competition is first best, and monopoly is not. The striking feature of this example is that introducing an indivisibility has no effect on the competitive equilibrium - but leads to an additional welfare loss under monopoly. That is - in this example -indivisibility strengthens rather than weakens the case against intellectual property.

We construct a specific case of an innovation chain with the desired properties. Specifically, suppose that for $\theta_1 < 0, \theta_2 > 0$ the period utility function is

$$u(c) = \begin{cases} -(1/\theta_1)\, c^{-\theta_1} & c \leq 1 \\ 2 - (1/\theta_2)\, c^{-\theta_2} & c > 1 \end{cases}$$

that is, it is an elastic CES below $c = 1$ and an inelastic CES above that consumption level. This satisfies the assumption of an asymptotically CES we used above in our competitive analysis of innovation chains. Suppose first that there is no indivisibility and no depreciation ($\zeta = 1$) and that the initial capital stock is $k_0^0 = 1$.

As before, define consumption productive capacity $P_t = \sum_{i=0}^{I_t} \gamma^i k_t^i$. Asymptotically, the competitive growth rate of P_t is given by

$$g = (\delta\rho\gamma)^{1/(1+\theta_2)}$$

and P_t grows over time provided that $\delta\rho\gamma > 1$. Assume this is the case. Then competitive equilibrium will give rise to sustained innovation and will continue to do so when there is positive depreciation and a small indivisibility.

Consider, by contrast, a monopolist who has the right not only to profit from sales of his product, but to control what is done with the product after it is sold. The utility function is designed so that the global maximum of revenue $u'(c)c$ takes place at a unit of consumption. The monopolist starts with a unit of capital that does not depreciate, so he can produce a unit of consumption each period. Because it is impossible to do better than this,

[10] Since 1962, the U.S. Congress has extended the term of copyright retroactively on each occasion that any existing copyright has been scheduled to expire. The U.S. Supreme Court recently ruled that the "limited times" envisaged in the U.S. Constitution means an infinite time.

this is the optimum for the monopolist, more or less regardless of modeling details of timing and commitment. The monopolist will not choose to innovate because any investment to do so must necessarily reduce current-period revenues below the maximum, while it cannot raise revenue in any future period. Similarly, the monopolist will not allow anyone else to innovate.

The point is a fairly simple one. Monopolists as a rule do not like to produce much output. Insofar as the benefit of an innovation is that it reduces the cost of producing additional units of output but not the cost of producing at the current level, it is not of great use to a monopolist. In this example, the monopolist does not innovate at all and output does not grow at all, while under competition, repeated innovations take place and output grows without bound.

Notice the significant role played in this example by the durability of the capital good (absence of depreciation). Other authors, such as Fishman and Rob [2000], have emphasized the role of durability in reducing the incentive of monopolists to innovate. Here the absence of depreciation is crucial because, without an indivisibility, the optimal method of replacing depreciated capital would be through innovation, even for a monopolist.

Introduce now into the model a small amount of depreciation, but still no indivisibility. The competitive equilibrium remains first best, and there is still a welfare loss from monopoly. However, as we just pointed out, the monopolist is as innovative as the competitive market, introducing a new type of good every period to cover depreciation.

Now we introduce a small indivisibility - the condition usually thought least conducive to competition. Again we have constructed the example so that competition still achieves the first best. However, the monopolist may cease to innovate in the presence of the indivisibility. Specifically, what is required is that the depreciation rate be small enough that the amount of capital required to invest to replace the depreciated old capital is less than the threshold for producing a single unit of new capital via the ρ technology.

This result should be underlined because it can be traced directly to the different incentives to innovate under the two market regimes. The competitive industry has an incentive to produce additional output that goes over and above the need for replacing the depreciated goods. As long as the consumer marginal valuation is high enough to cover the cost of production, a competitive industry will increase output as entrepreneurs try to maximize the overall size of the capital stock, and so is more likely to reach the threshold requirement at which innovation becomes possible. All this fails under monopoly. If the previous discussion reminds the reader of, for example, the telecommunication industry before and after the breakup of the national monopolies, the reader is quite correct.

Earlier in this section we singled out as particularly strong the assumption that the monopolist fully controls all kind of production that uses the new good, and can do this forever. It is therefore worth pausing a moment to consider if our conclusions rely too much on this strong assumptions. A little inspection shows they do not. Any form of binding monopolistic protection of the new good results in a welfare loss over the competitive legal environment in which intellectual property, via patents and copyrights, is prevented. Specifically - consider a T-period monopoly, followed by competition after that; this resembles current patent protection in most OECD countries. Since revenues for the monopolist strictly decline from the initial condition, the monopolist will suppress innovation for T periods - i.e. as long as he can. If subsequent intellectual property is awarded, say randomly by a patent race among the various possible producers of the new good (who we can reasonably take as all equal ex-ante), then the situation is even worse - innovations occur only every T periods instead of every period, and the monopolists who follow the first will actually allow the capital stock to depreciate (or even destroy it) because this strategy gets them closer to the revenue maximum. Alternatively, assume the monopolist only controls the β but not the ρ technology; in other words: only the monpolist can reproduce the new good i, but any of his customer may use her acquired share of the stock k_t^i to try to become a new monopolist by using the ρ activity. The presence of an indivisibility, once again, reinforces the socially damaging role of the monopolist. In principle, this would want to always keep the amount of k_t^i available in the market to the new potential innovators below the threshold level $\underline{h}$, so as to prevent the introduction of the new good. The higher $\underline{h}$ is, the easier this prevention becomes. Once again, the presence of an indivisibility weakens the case for intellectual monopoly and reinforces the view that competition can innovate at least as much, in fact: strictly more, than monopoly. The opposite of the received wisdom.

6. The New Economy and the Superstars

We turn now to a positive application of our theory of innovations and of their adoption. We use it to model the "economics of superstars". Next, we claim that our interpretation of superstars suggests that a very similar, and very simple, mechanism may be the underlying cause of the increase in skill premia in wages and earnings which has been widely observed during the last few decades.

The phenomenon of *superstardom* was defined by Rosen [1981, p. 845] as a situation "wherein relatively small numbers of people earn enormous amounts of money and dominate the activities in which they engage." Its

puzzling aspect derives from the fact that, more often than not, the perceivable extent to which a superstar is a better performer or produces a better good than the lesser members of the same trade is very tiny. Is superstardom due to some kind of monopoly power, and would it disappear in a competitive environment?[11] Our theory shows that when there are indivisibilities, technological advances in the reproduction of "information goods" may lead to superstardom, even under perfect competition. Hence, our model predicts that superstars should abound in industries where the main product is information which can be cheaply reproduced and distributed on a massive scale. Such is the case for the worlds of sport, entertainment, and arts and letters, which coincides with the penetrating observations (p. 845) that motivated Rosen's original contribution.

For simplicity, we consider a world in which all consumption takes place in a single period, but our results extend directly to an intertemporal environment. There are two kinds of consumption goods. The first is the information good we concentrate upon, while the second can be interpreted as a basket of all pre-existing goods. Specifically, we assume utility of the form $u(c)+m$, where c is the information good. There are two kinds of potential producers, A and B, each with a single unit of labor. The two producers are equally skilled at producing the second good: a unit of labor produces a unit of the second good. However, A produces information goods that are of a slightly higher quality than those produced by B. To be precise, we assume that one unit of type A labor can produce $(1+\varepsilon)\beta$ units of good c, while one unit of type B labor can produce β units of good c.

This case, without indivisibility, does not admit superstars, in the sense that the price of type A labor must be exactly $1+\varepsilon$ the price of type B labor. Since type A labor is more efficient at producing the information good, type B labor will be used in the information sector only after all type A labor is fully employed in that sector. Suppose that this is the case. Let ℓ_2 denote the amount of type B labor employed in the information sector. Then, the equilibrium condition is simply $\beta u'(\beta(1+\varepsilon)+\beta\ell_2)=1$. If $u'(c)$ is eventually inelastic, then ℓ_2 must fall as β rises, and producer B will be forced out of the information good market. However, with good 2 as numeraire, it will always be the case that B will earn 1 and A will earn $1+\varepsilon$.

With an indivisibility, however, the situation is quite different. Suppose that it costs a fixed amount C to operate in the information good market at all. When ℓ_2 falls below C producer B no longer finds it profitable to participate in the information goods market and drops out entirely. This occurs

[11] Our thanks to Buz Brock for suggesting that we look at this problem through the lens of our model, and to Ivan Werning for pointing out an embarrassing mistake in an earlier version of this paper.

when $\beta u'(\beta(1+\varepsilon)+\beta C)=1$. In this case producer B of course continues to earn 1. However, prices in the information goods market now jump to $\beta u'(\beta(1+\varepsilon))$, and producer A now earns $\beta u'(\beta(1+\varepsilon))(1+\varepsilon)$, which will be significantly larger than $1+\varepsilon$.

The argument can easily be generalized to a dynamic setting with capital accumulation, endogenous labor supply, and so forth. It shows quite starkly that, under very common circumstances, the simplest kind of technological progress may have a non monotone and non homogeneous impact on the wage rate of different kinds of labor. Our model predicts that continuing improvements in the technology for reproducing "information goods" have a non monotone impact on wages and income inequality among producers of such goods. Initially, technological improvements are beneficial to everybody and the real wage increases at a uniform rate for all types of labor. Eventually, though, further improvements in the reproduction technology lead to a "crowding out" of the least efficient workers. When the process is taken to its natural limit, this kind of technological change has a disproportionate effect on the best workers. For large values of β, the superstar captures the whole market and has earnings that are no longer proportionate to the quality of the good it produces or its skill differentials, which are only slightly better than average.

To an external observer the transition between the two regimes may suggest a momentous change in one or more of the underlying fundamentals. In particular, one may be lead to conclude that the observed change in the dynamics of skill premia is due either to a shift from neutral to "skill biased" technological progress, or to a dramatic variation in the relative supply of the two kinds of labor, or, finally, to large changes in the skill differentials of the two groups. These are the main interpretations that a large body of recent literature has advanced to understand the evolution of wages during the last twenty five years. While one or more of these explanations may well be relevant, our simple example shows it needs not be and, we would argue, it certainly is not for those sectors in which "information goods" are produced. We find the explanation outlined here not only simpler but also, plainly, more realistic.

Our point of view puts at the center stage the working of competitive forces when there is indivisibility and the unavoidable consequences of the law of comparative advantages. Our theory predicts that even very small skill differentials can be greatly magnified by the easiness with which information can be reproduced and distributed. It also predicts that the increased reproducibility of information will continue generating large income disparities among individuals of very similar skills and in a growing number of industries.

7. CONCLUSION

The danger of monopoly and the power of competition have been recognized by economists since Adam Smith. The particular dangers of government enforced monopoly are now well understood, and a substantial effort is underway to deregulate government enforced monopolies and allow competition to work for a large number of markets and products. Strangely, both the economic literature on technological innovation and growth and that on the optimal allocation of intellectual property rights have been immune to careful scrutiny from the perspective of competitive theorists.[12] During the last century, the myth that legally enforced monopoly rights are necessary for innovation has taken a strong hold both in academic circles and among distinguished opinion makers.[13] Hence, the widespread intellectual support for political agendas claiming that strong monopoly rights on intellectual and artistic products are essential for economic growth. Current research on innovative activity focuses on monopolistic markets in which fixed costs and unpriced spillovers (externalities) play center stage. Monopoly pricing of the products of human creativity is seen as a small evil when compared to the bounties brought about by the innovative effort of those same legally protected monopolies. The ongoing debate about the availability and pricing of AIDS drugs and other medicines is a dramatic case in point. The conflict over Napster, Gnutella and other tools for distributed file sharing is a less dramatic but equally significant example of such tension.

Our goal here has been to establish than when its functioning is carefully modeled, competition is a potent and socially beneficial mechanism even in markets for innovations and creative work. We have argued that the crucial features of innovative activity (large initial cost, small cost of reproduction) can be properly modeled by introducing a minimum size restriction in an otherwise standard model of activity analysis with constant returns. We have shown that the novel conclusions reached in this simple model are maintained and enhanced when a chain of innovations is considered. In this sense, our model is one of positive economics insofar as it explains what has

[12]Leaving aside our own work, the initial version of which circulated in 1997, we know of one other, partial, exception to this rule. Hellwig and Irmen [2001] embed in an infinite horizon general equilibrium context a model, originally due to Bester and Petrakis [1998], in which infinitesimal competitive firms face a fixed cost plus a strictly increasing marginal cost of production. In the appropriate circumstances, inframarginal rents are enough to compensate for the fixed cost, allowing for the existence of a competitive equilibrium. Once new goods are introduced, though, the knowledge embodied in them is again a nonrivalrous good. Hence, also in this case, the competitive equilibrium is suboptimal, because knowledge spillovers are not taken into account by innovators.

[13]A look at very recent issues of *The Economist* or of *Business Week* easily confirms this.

happened, happens, or would happen, in markets where innovative activity is not granted legal monopoly rights. Such markets have existed and thrived through most of history.[14] Markets for competitive innovations still exist and thrive in contemporary societies, insofar as most entrepreneurial activity is de facto not covered by legal monopoly protection. This is especially important for understanding developing countries, where the adoption by small and competing entrepreneurs of technologies and goods already used or produced in the most advanced countries are tantamount to competitive innovation. The viability of competitive innovations is also supported by an array of examples from the advanced countries. After Napster the market for recorded music has turned competitive with at most a modest reduction in the production of new music.[15]

We also stress the normative implications of our model. Showing that innovations are viable under competition should cast doubts on the view that copyrights and licensing restrictions are to be allowed for the sake of sustaining intellectual production. For products that are both in elastic demand and easily reproducible, our analysis shows that the right of first sale at competitive prices is more than likely to cover the sunk cost of creating a new good. This is even more so if one considers that, in many instances, the innovative entrepreneur is a natural monopolist until substitutes are introduced, an event that may take a significant amount of time. This should invite a reconsideration of the sense in which the current 20 years of patent protection serves any social purpose, beside that of increasing monopoly profits above the cost of R&D and providing distortionary incentives for socially wasteful patent races, defensive patenting, and other legal quarrels. Further, the analysis of innovation chains takes us beyond the traditional welfare triangle costs of monopoly, clarifying why the rent-seeking behaviors induced through government grants of monopoly are likely to hinder rather than promote innovation.

Among the many topics of research mentioned but left unsolved by this paper, one looms particularly large. Competitive behavior when indivisibilities are binding is very poorly understood. When competitive rents are insufficient to recover production costs, the situation becomes akin to a public goods problem: under competition it becomes necessary to collect payments in advance, contingent on the good being created. While a theory of general equilibrium with production indivisibility remains to

[14]Landes [1998] is a recent review containing abundant evidence of this.

[15]There is debate over how much of the reduction is due to "piracy" and how much to the recession. See Leibowitz [2002] who argues that the data suggest that the long-run impact of *de facto* elimination of copyright for music will result in about a 20% reduction of sales revenue for recorded music.

be fully worked out, the literature on public goods provides many clues. We should first distinguish between situations where there is competition among innovators and situations where there is a single innovator with a unique product. In the former case, for example, we have drug companies competing to develop well-defined products, such as a vaccine for AIDS. The current patent system awards, without charge, a monopoly to the first past the post. The problems with patent races are well documented in the literature, for example, in Fudenberg et al. [1983]. To this we would simply add the obvious fact that it is possible to have competitors for patents compete on dimensions other than the race to be first. It is possible, for example, to award patents to the inventor that promises the lowest licensing fees, conditional on products quality standards. The current patent system is akin to an auction in which the good is sold to the first bidder, rather than the highest bidder. While such a system has the advantage to the seller that it results in a quicker sale, we do not often see such systems used in the private sector. We suspect there may be a reason for that.[16]

Turning to the case of an innovator with a unique product, such an individual has a natural monopoly as the only person capable of providing the initial copy. The key issue is whether such a natural monopolist should also be awarded the right not to compete with his own customers as is the case under copyright and patent law and often enforced as well through contractual licensing provisions. The issue, in other words, is the social desirability of enforcing downstream licensing provisions for intellectual products. The obvious fact is that if the good would be produced in the absence of such licensing provisions, there is no benefit to enforcing them and doing so will generally lead to distortions, as in our example of innovation chains. As we have indicated, in many practical circumstances the indivisibility does not bind and downstream licensing provisions are undesirable. When the indivisibility does bind, disallowing downstream licensing leaves a situation similar to a public good problem with (some degree of) nonexcludability. Although there are some results on this class of problems, for example, Saijo and Yamato [1999], the theory of public goods with nonexcludability is still underdeveloped. However, it is by no means true that public goods cannot be provided voluntarily when there is a certain degree of nonexcludability. For example, if it is possible to identify a group of n consumers, each of whom values the good at least v, then it is clearly possible to raise nv, by committing to provide the good only if all n consumers each pay v.[17] In other words, competition can still function, even in the presence of indivisibility and in the absence of downstream licensing.

[16] Kremer [2000] contains a number of interesting ideas in this direction.

[17] See Boldrin and Levine [2002b] for a simple model.

The point we should emphasize most strongly is that, as an allocational mechanism, competition leads to inefficiency only insofar as it leads to particular goods not being produced when socially valuable. We have emphasized the ability of competitive markets to generate revenues under a variety of circumstances. As our example of the superstars points out, competitive rents when reproduction costs are low can be disproportionate to the cost of being "best" rather than "good" even in the absence of patent protection.

REFERENCES

[1] Acemoglu, D. and F. Zilibotti (1997), "Was Prometheus Unbound by Chance? Risk, Diversification, and Growth," *Journal of Political Economy* **105**, 709-751.

[2] Aghion, P. and P. Howitt (1992), "A Model of Growth through Creative Destruction," *Econometrica* **60**, 323-351.

[3] Arrow, K.J. (1962), "Economic Welfare and the Allocation of Resources for Invention,", in Richard Nelson (ed.), *The Rate and Direction of Inventive Activity*, Princeton, NJ: Princeton University Press.

[4] Bester, H. and E. Petrakis (1998), "Wage and Productivity Growth in a Competitive Industry", CEPR Discussion Paper 2031.

[5] Boldrin, M. and D. Levine (1997), "Growth under Perfect Competition," UCLA and Universidad Carlos III de Madrid, October.

[6] Boldrin, M. and D.K. Levine (2002a), "Factor Saving Innovation," *Journal of Economic Theory* **105**, 18-41.

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

[8] Boldrin, M. and D.K. Levine (2003), "IER Lawrence Klein Lecture. The Case Against Intellectual Monopoly," forthcoming in *The International Economic Review.*

[9] Dixit, A.K. and J.E. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity,", *American Economic Review* **67**, 297-308.

[10] Fishman, A. and R. Rob (2000), "Product Innovation by a Durable-Good Monopoly," *RAND Journal of Economics* **31**, 237-252.

[11] Fudenberg, D., R. Gilbert, J. Stiglitz, and J. Tirole (1983), "Preemption, Leapfrogging, and Competition in Patent Races," *European Economic Review* **22**, 3-31.

[12] Grossman, G.M. and E. Helpman (1991), "Quality Ladders in the Theory of Growth," *Review of Economic Studies* **58**, 43-61.

[13] Hellwig, M. and A. Irmen (2001), "Endogenous Technological Change in a Competitive Economy", *Journal of Economic Theory*. **101**, 1-39.

[14] Jones, L.E. and R.E. Manuelli (1990), "A Convex Model of Equilibrium Growth: Theory and Policy Implications," *Journal of Political Economy* **98**, 1008-1038.

[15] Kremer, M. (2000), "Creating Markets for New Vaccines. Part I: Rationale," National Bureau of Economic Research working paper No. 7716, May.

[16] Lamoreaux, N. and K. Sokoloff (2002), "Intermediaries in the U.S. Market for Technology: 1870-1920," NBER working paper 9017.

[17] Landes, D. S. (1998), *The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor*, New York: Norton.

[18] Leibowitz, S. (2002), "Record Sales, MP3 Downloads and the Annihilation Hypothesis," mimeo, University of Texas, Dallas.

[19] Lewis, M. (1989), *Liar's Poker*, New York: Norton.

[20] Lucas, E.R. Jr. (1988), "On the Mechanics of Economic Development," *Journal of Monetary Economics* **22**, 3-42.

[21] von Neumann, J. (1937), "Über ein ökonomisches Gleichungs-System und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes,", in K. Menger (ed.) *Ergebnisse eines Mathmatischen Kolloquiums*, No. 8. Translated as "A Model of Economic General Equilibrium," *Review of Economic Studies* **XIII** (1945-1946).

[22] Plant, A. (1934), "The Economic Aspect of Copyright in Books," *Economica*, 167-195.

[23] Quah, D. (2002), "24/7 Competitive Innovation," mimeo, London School of Economics.
[24] Rebelo, S. (1991), "Long-Run Policy Analysis and Long-Run Growth,", *Journal of Political Economy* **99**, 500-521.
[25] Romer, P.M. (1986), "Increasing Returns and Long Run Growth," *Journal of Political Economy* **94**, 1002-1037.
[26] Romer, P.M. (1990a), "Are Nonconvexities Important for Understanding Growth?" *The American Economic Review (Papers and Proceedings)*, **80**, 97-103.
[27] Romer, P.M. (1990b), "Endogenous Technological Change," *Journal of Political Economy* **98**, S71-S102.
[28] Rosen, S. (1981), "The Economics of Superstars," *The American Economic Review* **71**, 845-858.
[29] Saijo, T. and T. Yamato (1999): "A Voluntary Participation Game with a Non-Excludable Public Good," *Journal of Economic Theory* **84**, 227-242.
[30] Schroth, E. and H. Herrera (2001), "Profitable Innovation Without Patent Protection: The Case of Credit Derivatives," mimeo, New York University.
[31] Scotchmer, S. (1991), "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law," *Journal of Economic Perspectives*, **5**.
[32] Schumpeter, J. (1911), *The Theory of Economic Development*, English translation: New York: McGraw Hill, 1934.
[33] Shell, K. (1966), "Toward a Theory of Inventive Activity and Capital Accumulation," *The American Economic Review (Papers and Proceedings)*, **56**, 62-68.
[34] Shell, K. (1967), "A Model of Inventive Activity and Capital Accumulation," in *Essays on the Theory of Optimal Economic Growth* (K. Shell, ed.), Cambridge, Massachusetts: MIT Press, 67-85.
[35] Stigler, G. J. (1956), "Industrial Organization and Economic Practice," in *The State of the Social Sciences* edited by Leonard D. White, University of Chicago Press.
[36] Tofuno, P. (1989), "First Mover Advantages in Financial Innovation," *Journal of Financial Economics*, **3**, 350-370.
[37] Varnedoe, K. (1990), *A Fine Disregard*. New York: Abrams, Harry N.

UNIVERSITY OF MINNESOTA AND UCLA