Federal Reserve Bank of Minneapolis
Research Department Staff Report 339

June 2004

IER Lawrence Klein Lecture: The Case Against Intellectual Monopoly

Michele Boldrin*

Federal Reserve Bank of Minneapolis
and University of Minnesota

David K. Levine*

University of California, Los Angeles

ABSTRACT

In the modern theory of growth, monopoly plays a crucial role both as a cause and an effect of innovation. Innovative firms, it is argued, would have insufficient incentive to innovate should the prospect of monopoly power not be present. This theme of monopoly runs throughout the theory of growth, international trade, and industrial organization. We argue that monopoly is neither needed for, nor a necessary consequence of, innovation. In particular, intellectual property is not necessary for, and may hurt more than help, innovation and growth. We argue that, as a practical matter, it is more likely to hurt.

*Both authors are grateful to the National Science Foundation for financial support. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

The modern literature on economic growth focuses on technological innovation, its determinants and its impediments, as the key for understanding long-run economic development. A large portion of the modern industrial organization literature focuses on technological innovation as the driving force behind the evolution of firms and industries. Applied and theoretical literature in the field of international trade conceives trade as mostly due to product differentiation springing from technological innovation and the introduction of new goods. In these and other fields of economic analysis, innovation is both cause and effect of monopoly power. Cause, as the innovative firm is assumed to gain, at least some temporary monopoly following the introduction of the new product; and effect, as entrepreneurs would not undertake the innovative effort absent the perspective of earning future monopoly profits. This two-way link between innovation and monopoly power has become, since Schumpeter first advanced it in the late 1940s, a dominant doctrine in many fields of economic theory: monopoly profits are the necessary cause and the natural effect of innovative activity. A standardized model of technological innovation has become common currency among scholars working in most areas of economics; it is a model in which the innovation is the disembodied and non-rivalrous outcome of the initial investment by the entrepreneur. Because of the non-rivalrous nature of innovations this model predicts that, in the absence of legal enforcement of the monopoly power ascribed, via intellectual property, to the original creator, copies of the new product would be reproduced by everybody else at a negligible constant marginal cost, thereby leaving the innovator in the dust. Absent the monopoly that intellectual property creates, entrepreneurs would not bother to innovate. Hence, the key role played by intellectual property: no intellectual property, no innovation.

The irony is that while the “monopolistic” approach to innovation is widely regarded as a theoretical necessity, there is little empirical

evidence to support the crucial underlying assumption of increasing returns to scale. The goal of this paper is to argue that the standard competitive model provides a more solid foundation for the study of growth and innovation, and that there is no theoretical need of postulating either increasing returns or monopoly power to understand the dynamics of innovation. One consequence of improving our understanding of how innovation takes place in a competitive environment is that it better enables us to focus on some fundamental weaknesses in standard arguments for intellectual property. In fact intellectual property may be damaging for innovation, growth, and overall social welfare; the monopoly profits generated by intellectual property have played, and still play, a much more secondary role than is commonly believed in determining the rate and pace of economic progress.

Let us focus first on the issue of intellectual property. In the common parlance, "intellectual property" confounds two different rights. One is the "right of sale" given to producers of ideas. It consists of the right to sell the fruits of intellectual work, in whatever form they can be packaged, embodied, and transmitted. This is not controversial; software producers have the right to sell the software packages they make and distribute the same way that watchmakers have the right to sell the watches they make and distribute. The second right associated with the term "intellectual property" refers to the power of producers of ideas to control how their products are used. This second right is enforced by means of an ever increasing set of legal tools: patents, copyrights, non-disclosure agreements, shrink-wrap agreements, and so forth. It is permitted only to selected groups of producers: to software designers but not to fashion designers, to producers of medicines but not (until very recently) to producers of financial securities, to writers of books but not to creators of culinary recipes, to software-makers but not to watchmakers. This ability to control downstream usage, and particularly to avoid competing with one's own customers, provides favored producers of ideas with monopoly power. This we refer to as "intellectual monopoly," and it

is this we wish to challenge as either necessary for innovation to take place, or socially desirable.

Conventional wisdom in industrial organization acknowledges that intellectual property in the second sense leads to undesirable legal "intellectual monopoly." However, it generally argues that this might be a good thing. In Kahn (1962), for example, we find the statement, “This issue is not one of principle but of practical social engineering: how much protection [...] of what kind is required and worth paying for.” This might seem uncontroversial – how much might include none at all, for example. Imagine, however, that such a statement referred to protection from international trade – such a statement would be controversial indeed, and we think it ought to be equally controversial in the case of intellectual goods.

As we mentioned, there are several strands of the existing literature that argue in favor of intellectual monopoly. Lucas (1988) and Romer (1986) argue in a growth theory setting that there are unpriced “spillovers” from innovative ideas, so that innovators are not fully rewarded for their incremental contribution to aggregate productivity. One way out of the asserted spillover problem would be via a complicated set of taxes and subsidies; another is to allow for monopoly power and relative profits. Still in the context of growth theory, Aghion and Howitt (1992), Grossman and Helpman (1991), and Romer (1990) argue that new goods are brought about via a technology for which aggregate increasing returns are unavoidable; hence, monopoly power is necessary for innovations to take place. In industrial organization Gilbert and Shapiro (1990) and Gallini and Scotchmer (2002) examine the theory of optimal patents. The starting point of their analysis is to assume that there can be no innovation without patent protection. In the theory of international trade Krugman (1980), and others after him, have developed models in which trade is due not so much to comparative advantages but to the introduction of differentiated goods via an increasing returns technology. Again,

intellectual monopoly is a requirement for product differentiation, division of labor, and international trade to take place.

The intuitive backbone of the conventional argument, common to all of the aforementioned theories, runs along the following lines. Information and ideas are a “public good” in the sense that, once an idea is discovered or a piece of information revealed, it can be appropriated and used by an unlimited number of people. In the terminology that has become popular since the work of Romer, ideas are non-rivalrous goods. Hence the externalities, or unpriced spillovers, from ideas. Alternatively, the presumption that new ideas have a public good nature means that there is a near-zero marginal cost of reproducing and distributing them, implying, if it is costly to produce the original idea, that there is increasing returns to scale in innovation. Conventionally, fixed cost plus marginal cost pricing with constant marginal cost (in this case zero) implies that a competitive firm must lose money. So without monopoly there will be no output of new ideas, and the conventional conclusion is that intellectual monopoly is necessary for the production of ideas and the creation of new goods. In the words of Schumpeter (1943), “If one wants to induce firms to undertake R&D one must accept the creation of monopolies as a necessary evil.”

Notice the logical structure of the argument we have just summarized: the presence of monopoly power is a logical consequence of the nature of the technology through which innovations are generated. This technological assumption has both positive and normative implications. On the positive side, this theory argues that, to model and understand the process of technological change, competitive theory is useless: when you see an innovation taking place, look for the monopolistic feature supporting it. On the normative side, the same assumption implies that legal enforcement of intellectual monopoly is a necessary evil, without which we would not be able to harvest the fruits of intellectual creation; hence, the issue is one of how much intellectual monopoly we should have, that it must exist is granted. Before moving on

to the presentation of an alternative description of the technology for innovation we will spend a few lines considering the building blocks of the conventional argument.

Let us start with the idea of unpriced spillovers. While there are certainly informational spillovers as ideas move from person to person, it is hard to see why they should go unpriced. Little justification is ordinarily given for this assumption, but the most likely culprit would seem to be employees moving from firm to firm. However, as Gary Becker (1971) astutely observed, "Firms introducing innovations are alleged to be forced to share their knowledge with competitors through the bidding away of employees who are privy to their secrets. This may well be a common practice, but if employees benefit from access to salable information about secrets, they would be willing to work more cheaply than otherwise." Plenty of supporting evidence notwithstanding, from apprentices' wages to the practice of pricing the academic quality of a department into the salary of new assistant professors, Becker's observation seems to have gone unnoticed. The same goes for an even earlier observation made by George Stigler, according to whom, "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, 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" (Stigler (1956)).

In more recent times, there have been some authors, such as Leibowitz (1985), who did recognize that spillovers are generally priced,

but for the most part this assumption has gone unchallenged. The idea of unpriced spillovers seems to be justified largely by the notion of agglomeration – that often similar firms locate near each other to take advantage of these positive externalities. But notice that this would be the case even if spillovers were priced, provided that transactions costs are lowered by locating nearby. Certainly, evidence supporting the idea that large and unpriced spillovers take place among innovating firms is scarce at best – Ellison and Glaeser (1999), who provide the strongest case for such an assumption, find at best very weak evidence that agglomeration is due to spillovers. Most studies find even weaker or no evidence for the allegedly pervasive unpriced spillovers. Acemoglu and Angrist (2000), for example, estimate average-schooling externalities at the U.S. state-level and find no evidence for significant externalities. Ciccone and Peri (2002) examine local labor markets to test if productivity increases with the average human capital of the workforce in the area where firms are located; the data reject this hypothesis. Most anecdotal evidence about industrial agglomeration, from Silicon Valley to the greenhouses of Almeria, suggests that firms do price informational and technological spillovers into the wages of their employees.

All this evidence notwithstanding, the idea that unpriced spillovers from new ideas are large remains widely held. This is often justified by means of the apparently obvious "fact" that ideas are nonrivalrous. This idea stems from a basic confusion about the economic value of ideas. Ideas, in their abstract form, are certainly nonrivalrous – unfortunately in their abstract form ideas also have no economic value. I can use any mathematical or physical theorem without minimally affecting the ability of other people to use the same theorem. But, and here is the catch, in order *for me* to be able to use such a theorem, it is not enough that the theorem exists in some abstract form – I must have acquired actual knowledge of the theorem. From an economic perspective, it is not abstract ideas that count, but rather copies of ideas embodied in either human or physical capital. My copy of the fundamental theorem of

calculus as embodied in my knowledge and understanding of calculus is a distinct economic entity from your copy, leading a separate economic existence. It is a different economic commodity in the obvious sense that if you were to die, taking your copy of the fundamental theorem of calculus with you, it would in no way limit my ability to make use of my copy of the fundamental theorem of calculus. “Ideas” are non-rivalrous in the same sense that my drinking from my cup of coffee has no effect on your ability to drink from your cup. And I can even less take advantage of your copy of your idea without your permission than I can drink from your cup of coffee without your permission.

This latter point – that I cannot access your copy of an idea without your permission – is important, because it is closely connected to the fallacy that ideas somehow are communicated automatically and costlessly. To use the fundamental theorem of calculus I must spend resources to learn it, the same way that to use any productive skill one must spend resources to acquire it, and the same way that to use any capital good one must purchase it. While abstract ideas may be nonrivalrous and disembodied, productive ideas are always embodied in either people or objects, and are as rivalrous as any other capital good. In fact productive ideas are, as even accountants have managed to recognize, parts and pieces of the capital stock of a society; their acquisition costs resources, and their reproduction and transmission cost resources. People owning a productive idea can earn income by teaching it, or by selling the objects in which it is embodied, or any combination of the two. The fact that in some cases the cost of transmitting an idea may be just a fraction of what it took to discover it is a feature of the technology of innovation which should be appropriately modeled but which, in any case, cannot justify the extreme assumption that transmission of ideas is costless and productive ideas are nonrivalrous goods.

The theory of innovation we present and discuss here is grounded on these elementary observations. Insofar as new commodities differ from other commodities, it is neither on account of the fact that they are non-

rivalrous public goods, nor because they generate positive externalities in the form of unpriced spillovers. So there is certainly no necessary role for monopoly in the theory or practice of innovation. New commodities differ from other commodities because their introduction requires someone to develop a "prototype," which may be costly and indivisible, and because there are, as long as the adjective "new" is applicable, few people capable of producing them. Productive capacity for a new commodity is therefore severely limited relative to that for an "old" commodity, and the cost of acquiring such productive capacity may often be quite high. The standard competitive model, when commodities and technologies are carefully defined, provides a useful and accurate description of this situation. It leads to a dynamic model of innovation and adoption, which in many fundamental respects is orthogonal to the conventional one. In the extreme, intellectual monopoly is not only superfluous for, but also damaging to, technological progress and social welfare; competition and imitation are, instead, good for technological progress and social welfare. This change of perspective about the nature and the causes of economic innovations has far-reaching implications for industrial organization, growth theory, and the theory of international trade. After summarizing a simple but formal representation of competitive innovation, we will discuss some of the consequences.

There is another dimension of the public policy debate over intellectual property that needs to be highlighted. The traditional view not only overstates the need for intellectual property – it obscures the fact that government grants of monopoly encourage socially costly rent-seeking behavior. Most industrial organization and law and economics literature concerned with optimal patent rules seems to forget that there is a very strong downside to government provision of legal monopolies. Although generally recognized outside the arena of intellectual monopoly, little attention is paid to the problem of rent-seeking in the context of innovation. Yet while evidence of unpriced spillovers and increasing

returns is weak, evidence of rent-seeking is strong and dramatic. Some recent examples suffice to make the point.

- The Sonny Bono copyright extension increased the length of copyright by 20 years retroactively; economists widely agree that the retroactive part of the extension serves no possible economic purpose.
- The Digital Millennium Copyright Act, illegalizing a variety of activities because they might have an impact on copyright holders, has been widely documented to have had a stifling effect on certain types of academic research and on free speech.
- Efforts are currently under way to legally mandate computer hardware in order to reduce copying. There is a possibility for substantial economic harm from legislation of this type because the computer industry, which is to bear the costs, is roughly an order of magnitude greater in size than the media industry that is the beneficiary.
- There is a long tradition of using the patent system as a rent-seeking device. This is the case, for example, with submarine patents, which are filed, but intentionally delayed by many years through the filing of constant amendments. Because the patent is never granted, it is never made public, and the date at which the patent expires is determined by the time at which the "submarine surfaces." This allows holding a claim to an idea that is currently useless, but might have some use in the future. Keep everything secret until someone else actually innovates and (usually at some substantial expense) develops the idea into something practical. After a nice business has developed the submarine surfaces, and demands royalties from the unsuspecting innovator. Obviously these activities contribute nothing to social welfare, but do detract from the incentive to innovate – who knows what submarines are lurking nearby? A very recent case in point is the effort by SCO to claim royalties for the free software system Linux, based on extremely weak legal claims.

2. Ordinary Economics of Scarcity

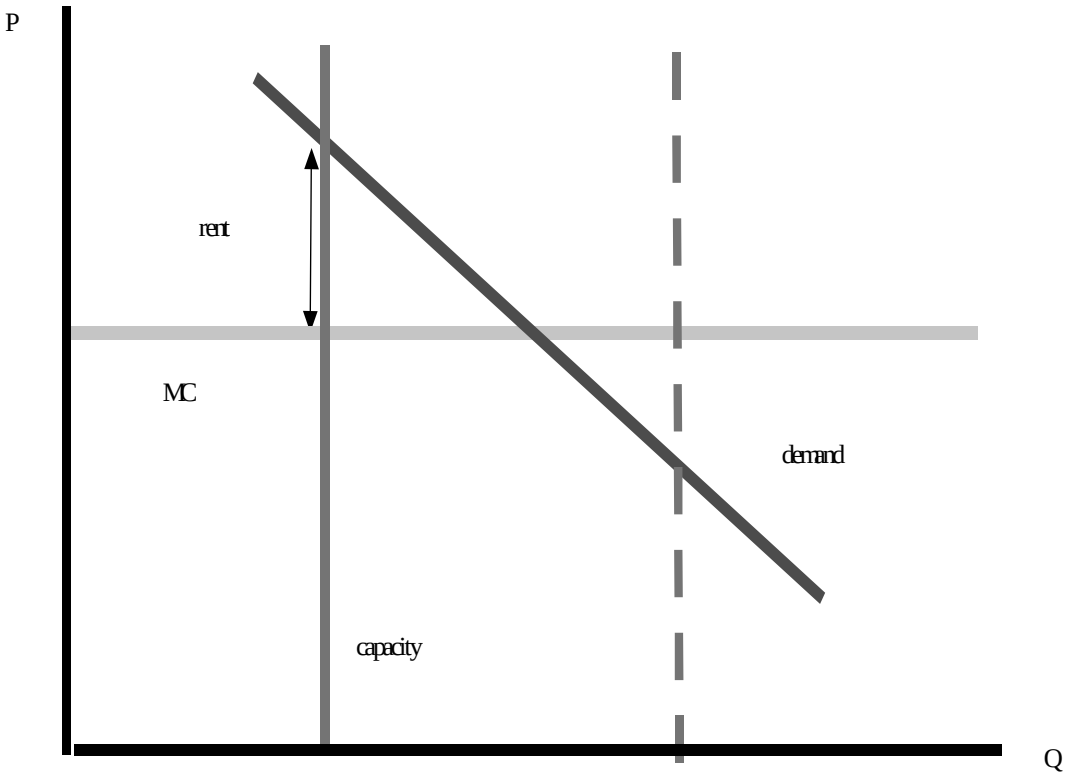
To understand where conventional reasoning about innovation goes astray, it is useful to discuss a simple example. For the sake of discussion, let us consider the creation of a new drug. We suppose (arbitrarily, but not ridiculously) that this drug takes a team of twelve expert biomedical researchers one year to invent. At the end of this year the team of twelve biomedical researchers is capable of producing the drug using tools and ingredients currently available on the market. The key point is that at the end of the year the knowledge is *embodied in the researchers* (and possibly some of the writing, machines, and materials generated by the R&D process they just completed) – no one can produce the drug unless the researchers tell them how to do it. For the new drug to be produced the team of researchers and their writings, machines, and materials are the stock of capital in which all useful knowledge is embodied – so far there is no unpriced spillover here.

Next we observe that it may be socially valuable to have other people know how to produce the drug, that is, to have more productive capacity of a useful commodity, rather than less. This is certainly the case whenever at full utilization of available capacity, the marginal utility of the last unit is higher than the marginal cost of producing it. For example, if a second team of twelve biomedical researchers knew how to produce the drug, they could set up a production line in Europe, while the original team was setting up production in the U.S., thereby satisfying the demand of many more people. As mentioned in the introduction, and despite the existing patent and copyright literature, it is a fact of life that transferring knowledge is a costly endeavor. How long would it take the biomedical team to explain to a group of inexpert economists how to produce the new drug? Given the huge literature on technology transfer, there is no mystery in the fact that it is costly to transfer productive knowledge. The mystery is: why do conventional economic theories of innovation ignore this fact?

If the second team is to learn how to make the new drug, there are two methods they can use to do it. First, they can “reinvent the wheel” by

simply replicating the efforts of the first team, spending a year doing research and obtaining the same stock of knowledge. Second, the first team can teach the second team how to do it. For the sake of concreteness, let us say that it takes one month to do it the second way (one month of team time for each of the two teams). The key observation is that the second method does not always dominate the first; the ranking depends on the relative price of the drug, the degree of impatience of consumers, and the opportunity cost of the biomedical researchers. The first method maximizes team time (two years to have two fully trained teams), but minimizes time until production can start (one year). The second method minimizes team time (one year and two months), but maximizes time until production can start (one year and one month). While team time has social value – so does beginning production one month earlier. Because beginning production one month earlier has social value, this immediately implies that the first team can sell its knowledge into a competitive market at a positive price, and not, as in the conventional story, at a price of zero. In fact, if it is socially optimal for the second team to produce the idea in parallel, then, since the first team can always recover the marginal social value of its knowledge, this price will necessarily cover the opportunity cost of having produced the knowledge in the first place. No government grants of monopoly are required to produce innovation in cases where it is socially optimal to have the initial knowledge produced by more than one team.

We now want to examine in greater detail what went wrong in the conventional story. If we consider the problem of building a shoe factory, we also face a constant marginal cost of producing shoes after the factory is built. Why is this not an issue? How can the fixed cost of the factory be covered? The answer is that shoe factories have a capacity constraint – if demand exceeds capacity then price will be above marginal cost, leading to competitive rents. This is illustrated in the diagram below.



In competitive equilibrium, of course, it will turn out that these rents exactly pay for building the shoe factory. Can we say in the case of ideas, as in the case of the shoe factory, that capacity is always chosen small enough that the competitive rent covers the cost of creation? In general, we cannot. With ideas the problem of indivisibility (or of minimum size) is significant. Indivisibility has some implications similar to that of fixed cost, but differs in important ways. In the example, there is no guarantee that the positive return is sufficient to compensate the research team for its time. It may be that the team would have to produce ¾ of an idea to be able to recover costs (better: that the productive capacity of a team of just nine researchers would be able to recover costs) – but this is not feasible because of indivisibility. This case is similar to the conventional story, and a legal monopoly on the new drug may be one way out of trouble. On the other hand, the social optimum might be such that saving a month in the start of production has social value exceeding a year of team time. In this

case, as we noted, the costs of the first team are necessarily covered by the competitive rent.

Which case would arise in practice will depend on the specific circumstances. While the traditional model predicts that monopoly power is *always* necessary for innovations to take place, the theory we are advancing does not claim that competitive rents are always enough to cover the discovery cost. In certain circumstances, when the initial indivisibility is particularly large relative to the demand for the new good, competitive arrangements would not do. But in most others, they will. The issue is therefore empirical, not one of principle. So that this does not seem a futile twisting of economic principles, a fact from Arnold Plant (1934): "During the nineteenth century anyone was free in the United States to reprint a foreign publication, 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]" where they did have copyright.

Our theory is consistent with facts such as this – highly innovative industries exist in which the law does not grant monopoly power to the innovator. This is an enormous puzzle for the standard theory. The existence of such “competitive innovation” directly contradicts conventional wisdom; from all kinds of design to investment banking, from advertising to civil engineering, to agricultural innovations (until the 1970s), competitive innovation is much more common than monopolistic innovation. Standard theory cannot explain this fact; our theory can, while at the same time accounting for the existence of monopolistic innovators.

3. One-Shot Innovation Under Competition

Ultimately, to understand whether an innovation will take place or not in a competitive environment, we must understand how much the new good/process is worth after it is created. To do this, one needs at least a

formal model for the equilibrium of a competitive industry after an innovation is introduced, to which we now move. General equilibrium and dynamic considerations will be introduced in the following section.

In this economy individuals live forever. There are many consumers, indexed by $c > 0$. In each period, consumers either consume one unit of the good, or not. The benefit to consumer c of consuming a unit of the good is $c^{-\psi}$, with $\psi > 0$. In other words, consumers are ordered by how they value the consumption flow of this good: consumers for whom c is small value it highly. Consumers also prefer to consume early rather than later: a unit of good consumed today is worth $\delta < 1$ of a unit of the same good consumed next period. In any period in which the good is not consumed, consumer c receives a payoff equal to zero, independently of how much he/she likes consumption.

Initially, there is a single prototype of the new commodity that generates the flow of consumption service. The inventor or producer owns this prototype. For concreteness, assume this is a durable good. Once sold, no downstream licensing or other kinds of monopolistic restrictions are possible. At each moment of time the prototype can either be used to generate a flow of consumption or reproduced. To make things less abstract, let us imagine the new good is a fresh recording of a new musical piece that is embodied in an MP3 file. Each copy takes one period to produce, and each MP3 that is copied produces $\beta > 1$ additional MP3's in that period. Our interpretation of a technology such as Napster or Audio Gnome is that it increases β, that is, it increases the number of MP3's that can be distributed (reproduced) to different consumers from a single master copy in a single time period. Note that there are two possibilities for the reproduction technology. With most goods, we assume that they are not simultaneously consumed and produced. This means that each consumer would face the decision of how much time to allocate listening to the MP3, and how much time reproducing it. However, at least in the case of MP3's, it may well be that it is possible to listen and copy at the same time. Since this 24/7 model of Quah [2002] is more favorable to

intellectual monopoly, as it makes copying less costly, we will adopt the 24/7 assumption that simultaneous copying and listening is possible.

Under competitive conditions, in the t-th period each MP3 sells for a single market price q_t. MP3's may also be rented for a single period for a rental rate r_t. Notice that consumers for whom $c^{-\psi} > r_t$ value the song more highly than the rental cost, and will choose to listen to an MP3 that period; consumers for whom $c^{-\psi} < r_t$ will choose not to listen to the MP3: if they have a copy, they prefer renting out their copy to someone else to listening to it themselves. Notice how in a competitive environment, everyone is potentially a buyer and a seller. We are interested in three primary questions. Is the price of the very first copy ever different from zero in such an environment? If it were, would it be enough to compensate the producer for its sunk cost? Finally, does the price of the first copy increase or decrease when new technologies increase β? Recall that, according to the standard model of innovation, the answers are, respectively, "no," "no," and "decreases." We will show that, in our theory, the answers are "yes," "it depends," and "increases."

According to standard competitive theory the sale price of an MP3 is just the present value of the rental rates. Since the rental rate is determined by the marginal consumer, it is $r_t = u'(c_t)$ where c_t is the number of MP3s consumed in period t. Note that since simultaneous consumption and production is allowed, production accumulates MP3s at a constant rate of β per period, and there was only one MP3 in period zero, the number of MP3s in period t is β^t. Hence, the price of MP3s at time τ is

$$q_\tau = \sum_{t=\tau}^{\infty} \delta^t u'(\beta^t)\beta^t.$$

In particular, q_0 is always a positive number. For finite values of β satisfying an obvious upper bound, it is also a finite number. Since q_0 is what the producer can earn from the first sale when he has no downstream protection at all (in practice he should be able to do better than this), there is money to be made for producers of intellectual products.

Is this competitive value of intellectual products enough to motivate the producers to spend the effort and time required? We do not know. To answer this question one needs to know the particular opportunity cost of time of the particular creator, which clearly varies from case to case.

We also want to understand the social impact of a technology which facilitates the reproduction of "idea-goods." Does it increase or decrease the value of intellectual products in a competitive market? Basically, received wisdom argues that cheap copying makes it impossible for innovators to earn back their production costs. If, in a competitive setting, increasing β lowered q_0 received wisdom would be correct – without downstream protection, less "idea-goods" would be created as a result of the advent of the new technology. What does happen to q_0 as the parameter β grows larger? The answer depends on ψ. If $\psi < 1$ demand is elastic. This is the empirically interesting case, at least when thinking at the early stages in the life-cycle of a new product. As β grows larger, it is easy to check from the equation above that the price of the very first unit, the one from which the innovator receives his payoff, increases. For the particular functional form we adopted q_0 actually goes to infinity as β approaches a finite value. Notice that, in all cases, the rate at which the price falls over time is proportional to β. Nevertheless with elastic demand and large β, the dramatic increase in the rate with which price falls over time is associated with a higher *initial* price and greater rent for the innovator.

In summary, under competition and in the empirically interesting case where demand is elastic, improving the technology for reproduction increases the first sale price. Contrary to assertions based on standard theory, careful inspection shows that the improved technology makes it much easier for a producer to recover sunk costs in a competitive market. This does not mean that the producer will argue against downstream licensing and in favor of increased competition: she will still be able to earn more revenue with a monopoly than under competition. But it is a

good argument for not giving in to the producer and granting them the monopoly: the social benefit of the monopoly (the ability to cover sunk costs and produce a socially desirable good) is reduced by the new technology. Indeed, in the case of music, the same computer technology that is increasing β with ambiguous consequences for q_0 is at the same time lowering the size of the indivisibility. The capabilities of a music studio that would have cost tens of millions of dollars several decades ago are now available using laptop computers and specialized software for thousands of dollars.

This establishes competitive markets as a viable institutional setting for fostering innovative activity. We move now to consider the general equilibrium implications of this approach in a growth theory context.

4. Competitive Innovation and Growth Theory

In this section we embed our theory of competitive innovation in a dynamic general equilibrium context. The main implications of our approach for growth theory and general equilibrium dynamics are well illustrated by an example of sequential innovation in which – despite the presence of an aggregate indivisibility – the patent system is strictly Pareto dominated by the absence of any intellectual monopoly. As a second implication of the theory of innovation under competition, we examine how the trade-off between introduction of new machines and accumulation of old machines leads both to cycles in innovation and a fully endogenous rate of growth.

4.1 Innovation and Welfare Theorems

We consider an economy in which an infinite number of different capital goods, indexed by $i = 0,1,2,\ldots$, can be introduced sequentially, with capital good i being a pre-condition for the introduction of capital good $i+1$. The stock of capital of quality i is denoted by k^i. Quality i capital may be used for several purposes. It may be used to produce $(\gamma)^i$ units of consumption, $\gamma > 1$, it may reproduce $\beta > 1$ units of itself, or it

can be used to produce $\rho < \beta$ units of the next quality capital $i+1$. We call these three alternative uses the γ, β, and ρ technology, respectively. Capital used in the γ technology depreciates at a rate ζ; capital used in the β and ρ technologies depreciates completely. Central to the idea of innovation is that there should be an indivisibility in the creation of new ideas, so we assume that the ρ technology is subject to an aggregate indivisibility of $\underline{k} > 0$; if less type i capital than $\underline{k}$ is used, then no output of the new capital good $i+1$ results. An amount $k > 0$ of capital good $i = 0$ is already available in period $t = 0$.

When capital is allocated in a feasible way among the three production technologies $k_t^i \geq k_t^{\gamma,i} + k_t^{\beta,i} + k_t^{\rho,i}$, output is given by

$$c_t = \gamma^i k_t^{\gamma,i},$$
$$k_{t+1}^i = \beta k_t^{\beta,i} + (1-\zeta) k_t^{\gamma,i}$$

and

$$k_{t+1}^{i+1} = \rho k_t^{\rho,i}, \text{ for } k_t^{\rho,i} \geq \underline{k}; \quad k_{t+1}^{i+1} = 0 \text{ otherwise.}$$

We also assume that technological change is socially desirable, so that $\rho\gamma > \beta$, and that growth is both feasible and desirable, so that $\delta\beta > 1$. This means, absent the indivisibility, that only the ρ technology would be used, never the β.

Denote with $p_t \geq 0$ and $q_t^i \geq 0$, respectively, the period zero present-value price of a unit of consumption and of a unit of capital of type i, available in period $t \geq 0$; we use consumption in period $t = 0$ as the numeraire. We are interested in the competitive equilibrium of such an economy, under the assumption that a complete sequence of markets is available at $t = 0$ for trade in all the dated commodities $\{c_t\}_{t=0}^{\infty}, \{k_t^i\}_{i,t=0}^{\infty}$.

There is a representative consumer, endowed with k units of k_0^0 at $t = 0$, and nothing after it. The period utility function $u(c_t)$ is strictly increasing, concave, and bounded below. The discount factor is $0 \leq \delta < 1$. We assume that the feasible present value of utility

$$U = \sum_{t=0}^{\infty} \delta^t u(c_t)$$

is bounded above. That is,

$$\sum_{t=0}^{\infty} \delta^t u[(\gamma\rho)^t] < \infty$$

holds. The problem of the consumer is then

$$Max_{\{c_t\}_{t=0}^{\infty}} \; U$$

subject to

$$\sum_{t=0}^{\infty} p_t c_t \leq q_0^0 k.$$

This is very standard, so we will not indulge discussing its properties.

Consider our economy at any point in time other than $t=0$. Denote with $x_t = (k_t^0, ..., k_t^i, ..., k_t^t)$ the vector of available capital stocks. Note that, at time t, no quality of capital $i > t$ can possibly have been introduced. Consider the maximization problem of a firm purchasing an input vector x_t in the current period and planning to sell its output vector y_{t+1} in period $t+1$. Depending upon which of the $(3\times t)$ activities is being used, the output vector may consist of any feasible combination of c_{t+1} and $(k_{t+1}^0, ..., k_{t+1}^i, ..., k_{t+1}^t, k_{t+1}^{t+1})$. The aggregate technology set is a closed and convex cone, pointed at the origin but truncated along one dimension, the one corresponding to the output of the new capital good k_{t+1}^{i+1}. The size of the firm is therefore indeterminate, and we may as well assume that only one representative, price-taking firm is in place. In an abstract setting the truncation of the aggregate technology set along the k_{t+1}^{i+1} dimension generates difficulties for the standard proof of existence of equilibrium, and for proving the second welfare theorem as well. While not insurmountable, addressing these technical aspects in the present context would take us too far astray from our main concern. We therefore proceed by temporarily adopting the simplification that $\underline{k} = 0$.[1] Setting $\underline{k} = 0$ eliminates the indivisibility constraint, fully restoring the convexity

[1] The case in which the indivisibility is binding will be addressed later in this section. The reader should consult Boldrin and Levine [2003] if interested in a detailed treatment.

of the production cone. The problem of a firm operating in any of the three sectors β, γ, and ρ is then rather standard. Given $\{p_t, q_t^i\}_{i,t=0}^{\infty}$ maximize period by period profits π_t. Profits are the difference between the present values of y_{t+1} and x_t. Given prices, firms compute their production plans $\{c_t\}_{t=0}^{\infty}, \{k_t^i\}_{i,t=0}^{\infty}$. Notice that in our model the period-technology set Ω_t, which is composed of all the mutually compatible combinations of input and output pairs (x_t, y_{t+1}), is state-dependent, and may change from one period to the next. It does change whenever prices $\left(p_t, q_t^i\right)$ and $\left(p_{t+1}, q_{t+1}^i, q_{t+1}^{i+1}\right)$ in two adjacent periods make it profitable to use part of the stock of capital k_t^i to introduce the new kind k_{t+1}^{i+1} a period later. The latter will be true whenever $q_{t+1}^{i+1} / q_{t+1}^i \geq \beta / \rho$ holds. This is the formal requirement for a *perfectly competitive innovation* to take place.

Given the pair $(q_0^0, \{p_t\}_{t=0}^{\infty})$ let $\tilde{c} = (\tilde{c}_0, ..., \tilde{c}_t, ...)$ be the unique sequence maximizing $U(c)$. Define a *competitive equilibrium* for this economy as a collection of sequences

$$\left(\{\tilde{c}_t\}_{t=0}^{\infty}, \left\{\{k_t^{i,j}\}_{i,t=0}^{\infty}\right\}_{j=\gamma,\beta,\rho}\right)$$

for quantities, and

$$\left(\{p_t\}_{t=0}^{\infty}, \{q_t^i\}_{i,t=0}^{\infty}\right)$$

for prices such that, given

$$\left(\{p_t\}_{t=0}^{\infty}, \{q_t^i\}_{i,t=0}^{\infty}\right), \text{ the sequence} \left(\{\tilde{c}_t\}_{t=0}^{\infty}\right)$$

maximizes U, and the pair

$$\left(\{\tilde{c}_t\}_{t=0}^{\infty}, \left\{\{k_t^{i,j}\}_{i,t=0}^{\infty}\right\}_{j=\gamma,\beta,\rho}\right)$$

maximizes profits, with $k_t^i \geq k_t^{i,\gamma} + k_t^{i,\beta} + k_t^{i,\rho}$ for all $t = 0, 1, \ldots$. Under constant returns and competition firms cannot make positive profits; hence, the sequence of equilibrium prices $(\{p_t\}_{t=0}^{\infty}, \{q_t^i\}_{i,t=0}^{\infty})$ must satisfy

$$p_t \gamma^i \leq q_t^i,$$
$$q_{t+1}^i \beta \leq q_t^i$$

$$q_{t+1}^{i+1}\rho \le q_t^i,$$

with equality whenever the relative technology (γ, β, or ρ) is used at a positive level. As $\underline{k} = 0$ and $\rho\gamma > \beta$ hold, it follows that, in equilibrium, $i = t$ and a new kind of capital is introduced in each period. Hence, the equilibrium behavior of our simplified model is equivalent to one with an endogenous capital ladder or with an exogenous vintage capital structure. This competitive equilibrium is obviously a Pareto Optimum, so that the first welfare theorem is satisfied. The second welfare theorem is also satisfied as the unique solution to the planner problem

$$V(k) = Max_{c=\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \delta^t u(c_t)$$

subject to

$$c_t = \gamma^i k_t^{\gamma,i},$$

$$k_{t+1}^i = \beta k_t^{\beta,i} + (1-\zeta)k_t^{\gamma,i},$$

$$k_{t+1}^{i+1} = \rho k_t^{\rho,i},$$

and $$k_t^i \ge k_t^{\gamma,i} + k_t^{\beta,i} + k_t^{\rho,i};\ k_0^0 = k \text{ given},$$

can be decentralized by prices $(\{p_t\}_{t=0}^{\infty}, \{q_t^i\}_{i,t=0}^{\infty})$ satisfying the conditions given earlier.

4.2 Aggregate Indivisibilities and Patterns of Innovation

Now we bring back the aggregate indivisibility $\underline{k} > 0$ and ask if this alters, and how, any of our previous results. It does, and along three directions. First, the Second Welfare theorem may fail: there exist optimal allocations that are not competitive equilibria, at least for the standard definition of competitive equilibrium adopted here. Second, the kind of competitive equilibrium we have defined earlier may fail to exist. Third, initial conditions matter, and may even affect the long-run rate of product innovation insofar as competitive equilibria differ from the simple vintage-capital-like pattern described above.

The intuition for why the Second Welfare Theorem may fail is straightforward, and certainly familiar to students of general equilibrium with indivisibilities. Suppose that there is 100% depreciation, and that

$V(k_t)$ denotes the social value function of the available stock of capital k_t. Notice that, in general, this may be a vector listing all kinds of capital available at a certain point in time. Assume

$$\rho \frac{\partial V(k_1^1 = 0)}{\partial k_1^1} > \beta \frac{\partial V(k_1^1 = 0)}{\partial k_1^0}. \tag{4.1}$$

Then, when $k_1^1 = 0$ it is profitable to introduce type 1 capital at the prevailing competitive prices; notice that inequality (4.1) is a necessary condition for it to be also socially beneficial to do so. On the other hand if

$$\rho \frac{\partial V(k_1^1 = \rho \underline{k})}{\partial k_1^1} < \beta \frac{\partial V(k_1^1 = \rho \underline{k})}{\partial k_1^0} \tag{4.2}$$

it is unprofitable to have built the minimum quantity of type 1 capital at the competitive prices that prevail after it is built. Notice that this may or may not be true when it is socially beneficial to introduce type 1 capital. Hence, if it is socially beneficial to introduce type 1 capital and (4.1) and (4.2) simultaneously hold, the efficient allocation cannot be decentralized as a competitive equilibrium.

In fact, these are exactly the same circumstances in which a competitive equilibrium may fail to exist. This is easy to see, as when no type 1 capital is introduced it is profitable to do so; but introducing type 1 capital cannot be made consistent with competitive equilibrium, since doing it would involve negative profits. So this example is a failure of both the second welfare theorem, and existence of a competitive equilibrium.

Both lack of existence and failure of the second welfare theorem have the same causes: an indivisibility $\underline{k}$ which is too large relative to the available stock of capital k_0^0, a vastly superior new technology ($\gamma\rho >> \beta$), and a rapidly decreasing marginal utility of consumption. Under these circumstances, but *only under these circumstances*, perfectly competitive innovation systems fail to deliver the social optimum, and lack a

competitive equilibrium in the usual sense.[2] Under these same circumstances, but *only under these circumstances*, intellectual monopoly may support a socially better allocation than competition can achieve. Notice though that, for this to be the case, we must assume that the monopolist can install capacity equal to $\underline{k}$ and then manage to produce strictly less than that, or price-discriminate among heterogeneous consumers.

If we extend the notion of competitive equilibrium to allow the case in which (4.1) and (4.2) hold, but the new capital good is not produced while the old one is still accumulated, something nearly optimal may be implemented.[3] Roughly speaking, this corresponds to a path along which capital of type $i=1$ is not introduced right away, as requested by the social optimum, but a few periods later. How many periods later will depend on the size of β. When the economy grows sufficiently, the first innovation occurs and the economy switches to the new kind of capital via the ρ technology. As the length of time to innovation is a decreasing function of β, the welfare distance between the competitive equilibrium and the social optimum is smaller the higher is the value of β in relation to ρ.

This example also shows that initial conditions matter in the competitive theory of innovation, which is the third difference between the model with and without indivisibilities. Assume the allocation just described, in which innovation is postponed for a few periods until enough old capital is accumulated via the β technology, is in fact a competitive equilibrium. Consider what may happen after the first innovation has taken place, say at time $\tau > 1$, and capital of type 1 has been introduced. Different "continuation paths" are possible, depending on the relative sizes of k_τ^1, $\delta\rho$, and $\underline{k}$. If the latter is the same for all $i = 0,1,\ldots$, as we assumed

[2] Other, possibly more relevant, notions of competitive equilibrium may exist under such circumstances, which also implement the social optimum.

[3] This is closely related to the equilibrium concept used in Acemoglu and Zilibotti (1997) to study the role of diversification in innovation, and which is based on Hart (1979) and Makowski (1980). See Boldrin and Levine (2003) for a more careful discussion.

so far, and $\delta\rho > 1$, then it is obvious that, after the constraint represented by the low level of the initial stock of capital has been relaxed by the τ periods of β-driven accumulation, the indivisibility will no longer matter. In this simple case, initial conditions only affect equilibria during the first few periods, and accumulation paths become identical in the long run; a new capital good is introduced in each period and consumption grows at the socially optimal rate independently from initial conditions.

Consider, although in passing, the case in which either the innovation technology is not very productive ($\delta\rho < 1$) so first best investment declines over time, or the indivisibility grows with the new kind of capital being introduced. In such circumstances a binding indivisibility is likely to come back and haunt the innovation process over and over again, no matter how high productive capacity has become. When this happens initial conditions matter also in the long run as they determine how many periods of β-driven accumulation are needed between one innovation and the next. Patterns of innovation are then cyclical, with the innovation's phase followed by more or less long periods of steady accumulation of the same kind of capital stock to be interrupted by further bursts of innovation, and so on.

Finally, we observe that these innovation cycles need not be driven by a physical indivisibility per se. There can be a similar cycle between capital widening and capital deepening when technological change is not neutral. Boldrin and Levine (2002) examine, without indivisibility, the case in which innovation is "factor saving" and is, therefore, intrinsically biased to reduce labor input per unit of output. While, in the simple indivisibility example above, both productivity and consumption grow faster during the innovation periods and nothing can be said about the employment level, in the factor saving model consumption grows slower or does not grow at all when a technological innovation is taking place, and employment decreases while productivity increases during that stage of the economic cycle.

Unlike the simple indivisibility example, the factor saving model generates movements in aggregate employment and productivity that are closer to observed one. In particular, a positive correlation emerges between the growth rates of total factor (or labor) productivity, employment, as well as a measures of investment in new capital. Many authors have called this the process of "creative destruction" and explained its appearance with a widespread presence of external effects and market power. However, it appears that competitive factor saving innovations – even without indivisibility – are sufficient for a reduction in employment (relative to trend) to generally accompany an increase in productivity (relative to trend). This casts doubts on recent claims that VAR estimates linking productivity growth to a reduction in employment constitute sufficient evidence for rejecting the idea that productivity shocks may account for a large portion of business cycle fluctuations. Contrary to such claims, the factor saving innovation model shows that the central prediction of a competitive model of endogenous innovation is, exactly, that above average productivity growth should come together below trend employment growth.

Besides application to productivity growth, our framework has important implications for the past and future of intellectual property law. Generally speaking, capital accumulation reduces the significance of indivisibilities. Consequently, it reduces the need for intellectual monopoly. As we have shown, the larger are β and ρ, the sooner the indivisibility becomes irrelevant, and innovations flow undeterred by competitive pricing. When technological change increases either β or ρ, intellectual monopoly becomes more, not less, socially wasteful relative to competition. So, of course, do innovations that reduce the size of the indivisibility.

The logic underlying competitive innovations in general equilibrium is essentially the same we exposed, in a partial equilibrium context, in the previous section. When initial productive capacity of the new good is small, competitive rents can be very large. As long as the

rents accruing to the innovator (better: innovators, as a continuum of identical households are acting here) are large enough to compensate for the opportunity cost of the $\gamma^i \underline{k}$ units of current consumption the innovation requires, competitive innovation takes place. When those rents cannot satisfy such a requirement, innovation is postponed for a number of periods. During such periods capital of the old kind is accumulated using the β technology, until it reaches a stock so large that the opportunity cost of $\gamma^i \underline{k}$ units of foregone consumption is small enough to be paid for by the competitive rents. Our model not only explains how and why competitive innovations take place, but also how and why they sometime do not take place, and, finally, it explains how and why they follow irregular cyclical patterns.

Finally, the reader must have noticed that nothing argued so far hinges on the use of linear production functions. Had we written $c_t = \gamma(k_t^{\gamma,i})$, $k_{t+1}^i = \beta(k_t^{\beta,i})$, and $k_{t+1}^{i+1} = \rho(k_t^{\rho,i})$, with $\gamma(\cdot)$, $\beta(\cdot)$, and $\rho(\cdot)$ increasing and strictly concave functions, every single result would have gone through. In fact, when strictly concave production functions are adopted, the theory predicts that old kinds of capital are not immediately discarded every time a new and more productive one is adopted. They are, instead, slowly phased out as the stock of the new capital is progressively built up, as it appears to be the case in reality.

4.3 The Social Cost of Intellectual Monopoly

In the absence of intellectual monopoly, we still expect that innovators will earn rents on their unique ideas. But, we have shown, in some circumstances this is no guarantee that the rent will be sufficient to cover the cost of innovating. However much an innovator can earn without a monopoly, surely an innovator can earn no less and perhaps sometimes more, with a government grant of monopoly. Surely, then, a patent or copyright system will result in more innovation than in its absence. Even granting the dubious proposition that government grants of monopoly have no social cost, this need not be the case. While each individual

innovator may have no less incentive to innovate with an intellectual monopoly, because each innovation will generally incorporate the innovations of earlier creators, the monopoly power of the latter will reduce the incentive to innovate faced by the former. Indeed, in the extreme case, there may be no innovation at all in the presence of intellectual monopoly. Our next goal is to develop this idea.

Because innovations generally build on existing ideas, that is, on earlier innovations – it is generally recognized in the economics literature that intellectual monopoly has an undesirable effect on future innovation. This is central to Scotchmer (1991), for example. The fact that less innovation may result with intellectual monopoly than without is highlighted in Boldrin and Levine (1999), from which we adopt the following example.

Consider the same technology and commodity space as in the previous section. To differentiate between monopoly and competition, we make the usual industrial organization assumption that demand is initially elastic, and eventually inelastic. More precisely we assume that for some $\theta_1 < 0, \theta_2 > 0$ period utility function is

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

so that it is an elastic CES below $c = 1$, and an inelastic CES above. This utility function is designed so that the global maximum of revenue $u'(c)c$ takes place at $c = 1$.

Finally, we assume that the economy is productive enough that, under competitive conditions, the indivisibility never binds. In the absence of intellectual monopoly, this implies that the first best and competitive equilibrium allocation has consumption and investment growing over time and that a new type of capital is introduced each period. Since investment is growing over time anyway, it follows that if $\underline{k}$ is sufficiently small the constraint will not bind, and the competitive equilibrium remains the same

with or without indivisibility: repeated innovations take place because rents are high enough to provide an incentive for innovators to undertake innovative activity.

Consider, for simplicity and by way of contrast, a patent system in which a complete monopoly is granted the patent holder forever. Suppose in fact, to simplify exposition, that the initial capital stock is $k_0^0 = 1$ and the monopolist starts with a unit of capital that does not depreciate. It can then produce a unit of consumption each period without any need for investment, as the whole stock is allocated to producing the consumption good, i.e. as $k_t^{\gamma,0}$, and remains fixed at one because of no depreciation. In other words, the innovator begins at the revenue maximum. This is monopoly heaven: the innovator simply sits tight and collects the money. Indeed, this is true more or less regardless of modeling details about timing, preferences, depreciation, and commitment about prices of future output. The only reason for innovation would be to achieve higher levels of output, and that would lower his profits, so the innovator never makes a second innovation. Not only that, but no one else can innovate without a license from the original innovator, so he prevents anyone else from innovating for the same reason. In contrast to the competitive equilibrium of thriving growth and innovation, here the patent system leads to the complete absence of innovation and total stagnation.

This example may seem somewhat contrived, since it is the absence of depreciation that eliminates any incentive to innovate. If instead the depreciation rate is small and we still have $\underline{k} = 0$, there is aggregate stagnation as the innovator maintains the level of consumption at one, but there is constant innovation as new kinds of capital are introduced in order to replace old depreciated capital. In this case, there is no less innovation (but there is less welfare) under monopoly than under competition. However, this assumes that the indivisibility is not a problem. Suppose instead that the indivisibility matters, so that $\underline{k} > 0$ – then, for any fixed level of the indivisibility, if the depreciation rate is small enough, the monopolist will again choose not to innovate at all as in

the first example. The logic is, intuitively, the same one we explored in section 4.2 for the case in which the second welfare theorem was not satisfied. When the productive capacity to be replaced is small, because it depreciates slowly, there is no reason to pay the high cost associated with the indivisibility even if $\rho\gamma > \beta$; one maximizes profits by replacing the small amount of depreciated capacity with old capital goods produced via the β technology. To anyone recalling the phone systems of the world during the great days of the telephone monopolies, this story must sound vaguely familiar. Finally, relax the patent system and allow the monopolist to control the β but not the ρ technology. Also in this case, in the presence of an indivisibility, the monopolist may be able to keep potential competitors out of the market. To achieve this, he simply needs to produce an amount k_t^0 low enough to render the cost of the minimum plant size ($q_t^0 \underline{k}$) unprofitable for the potential innovator. Recall that the latter needs at least $\underline{k}$ units of capital of type $i = 0$ to introduce capital of type $i = 1$, and that the market price q_t^0 of type $i = 0$ capital can be manipulated by the monopolist.

In short, in this world patent protection leads to strictly less innovation, and an indivisibility in the production of new goods makes the problem *worse*, and is an argument *against* patents, rather than in their favor.

4.4 Implications for International Trade Theory

The standard model of innovation plays a crucial role also in many theories of international trade and, in particular, in theories aiming to connect technological progress to trade, and growth. In the currently standard model of international commerce, trade takes place because monopolistic producers of intermediate capital goods (or of different varieties of consumption goods), who are scattered more or less randomly across countries depending on initial conditions, ship their products around to allow final output (or utility) to be produced in each country via a Dixit-Stiglitz production function (or utility function). The producers of

such goods are the innovators; they face an increasing returns technology exactly like the one we discussed, and criticized, in Sections 2 and 3. A particularly important feature of this model is that, as innovation increases the number of intermediate goods (varieties of consumption good) trade increases in step with output (utility). This feature yields two fundamental predictions, which appear to be consistent with empirical observations. The first, that trade in capital and intermediate goods becomes increasingly more important as innovations expand the number of available goods, and labor productivity grows. The second, that the growth rate of output, the division of labor, and market size are positively related.

Are increasing returns to scale and monopoly power needed to explain these facts? Adam Smith predicted much the same thing, absent, however, any claims of monopoly power:

> "As the accumulation of stock must, in the nature of things, be previous to the division of labor, so labor can be more subdivided in proportion only as stock is previously more and more accumulated. … As the division of labor advances, therefore, in order to give constant employment to an equal number of workmen, an equal stock of provisions, and a greater stock of material and tools than what would have been necessary in a ruder state of things, must be accumulated beforehand. The quantity of industry, therefore, not only increases in every country with the increase of the stock which employs it, but in consequence of that increase, the same quantity of industry produces a much greater quantity of work" [*Wealth of Nations,* Book II, 3-4]. "The increase of demand, besides, though in the beginning it may sometimes raise the price of goods, never fails to lower it in the long run. It encourages production, and thereby increases the competition of the producers, who, in order to undersell one another, have recourse to new divisions of labor and new improvements of art,

which might never otherwise have been thought of" [*Wealth of Nations*, Book V.i.e, 26].

Consider in our framework, a world with two countries, A and B, two ladders $i = 0,1,...$ and $j = 0,1,...$ of capital goods, and two consumption goods, c^1 and c^2. For each ladder, assume a set of production functions similar to those introduced in Section 4.1 . Let c^1 be producible via capital goods belonging to ladder $i = 0,1,...$, and c^2 be producible from capital good of ladder $j = 0,1,...$. In each country, introduce a representative consumer with utility function

$$U = \sum_{t=0}^{\infty} \delta^t u(c_t^1, c_t^2),$$

and endowed with initial stocks

$$(k_0^{A,i=0}, k_0^{A,j=0}) \text{ and } (k_0^{B,i=0}, k_0^{B,j=0}),$$

respectively. Consider first the two countries under autarky. Assume the indivisibilities $\underline{k}^i$ and $\underline{k}^j$ are large enough, relative to the initial endowments, to render socially undesirable, for each individual country, the introduction of new capital goods, either of ladder i or of ladder j. Then both countries will use their β technologies to increase the initial stock of capital of both types $i = j = 0$. For many periods neither country innovates, therefore growing at a slower rate than otherwise desirable.

Look next at the impact that opening up trade between the two countries may have on their rates of innovation adoption and productivity growth. Let the two countries be slightly asymmetric in their ability to innovate. To fix ideas, country A has a slight advantage in ladder i while B has an advantage in ladder j, i.e.,

$$\frac{\rho^{A,i}}{\rho^{A,j}} > \frac{\rho^{B,i}}{\rho^{B,j}}. \qquad (4.3)$$

Opening trade has the immediate effect of pooling sectoral demands and, when the two capital goods are tradable, it leads also to a pooling of sectoral resources. Even when the two capital goods are not internationally

mobile, trade doubles the size of demand for each consumption good, facilitating the specialization of each country in the ladder in which it has a comparative advantage. The increase in demand induced by international trade reduces the chances that an inequality such as (4.2) holds, thereby weakening the indivisibility constraint and making it much more likely that $k^{i=1}$ (respectively, $k^{j=1}$) is introduced in country A (respectively, country B) in period $t=1$. This increases welfare in both countries. The ideal situation is, obviously, the one in which not only consumption but also capital, from both ladders, is a tradable good. In this case the "demand pooling" effect is reinforced by a "resource pooling" effect. After trade begins, one expects country A to import part or all of $k_0^{B,i=0}$ from country B, with the latter doing the same with respect to $k_0^{A,j=0}$. Either way, the indivisibility constraints are weakened and the length of time it takes to innovate reduced, thereby unambiguously increasing welfare in both countries. Further, after trade is allowed the allocation of production between the two countries is completely determined by comparative advantages, not by increasing returns. In this version of the model, in which comparative advantages are purely technological as defined in (4.3) above, the initial distribution of the two stocks of capital has no impact on the patterns of specialization. Introducing transportation costs for shipping capital from one country to another will make the patterns of specialization dependent upon initial conditions, rendering the model more versatile and interesting, without affecting the other main results. Trade allows comparative advantages to play their traditional role, and higher rates of productivity growth are related to larger trade flows.

In summary, trade has three effects. It increases demand and, possibly, the amount of resources available, in country A and in country B to overcome the indivisibilities $\underline{k}^i$ and $\underline{k}^j$. It leads to an increase in specialization and in the international division of labor. Finally, it increases the growth rate of productivity and income in both countries. This is what Adam Smith predicted, and we have witnessed since.

5. Conclusion

The theoretical idea of this paper – that intellectual monopoly can lead to less rather than more innovation while competition can lead to more, and more efficient, innovation – is well illustrated through the story of James Watt. In most histories, James Watt is a heroic inventor, responsible for the beginning of the industrial revolution. But an examination of the facts suggests otherwise – while Watt is certainly a clever inventor who managed to get one step ahead of the pack, he remained ahead not through superior innovation, but by clever exploitation of the legal system. The fact that his business partner was a wealthy man with strong connections in Parliament was not a minor help.

Watt's significant invention, of the steam condenser, occurred while he was working on an older Newcomen steam engine in 1764. He worked intensively for six months building a model. After a series of improvements, Watt attempted to patent the idea in 1768, spending about the same amount of time doing so that he originally spent building his first model engine. In 1775, supported by his business partner Boulton, Watt secured an Act of Parliament extending his 1769 patent until the year 1800. Burke spoke eloquently in Parliament in the name of economic freedom and against the creation of unnecessary monopoly – but to no avail. Boulton's connections in Parliament were too solid to be defeated by simple principles. In 1782, Watt secured a further patent apparently in an effort to preempt his rival Wasborough, who beat him to the invention of the crank motion. More dramatically, in 1781, when the superior and independently designed Hornblower machine was first produced, Boulton and Watt went after him with the full force of the legal system – bankrupting and ruining Jonathan Hornblower in the process.

The effect of Watt on steam engine innovation is reflected in production. Prior to Watt, there were 130 steam engines in the U.K., mostly of the old Newcomen design. They were used primarily for pumping water out of mines. By 1800, when Watt's patents expired, there were at most 1000 steam engines used in the U.K. of which only 321 were

the superior Boulton and Watt engines, with the remainder being the older Newcomen engines. Fifteen years later, it is estimated that 210,000 horsepower is installed in England alone. It is only after the expiration of the Watt patents in 1800 that there is an explosion not only in the production of steam engines, but in steam engine innovation. New innovation in steam engines greatly increased the variety of applications, and in the next 30 years steam power finally came into its own as the driving force of the industrial revolution through the advent of the steam train, steamboat and steam jenny. Between 1800 and 1804 the most significant improvements, those of William Bull, Richard Trevithick, and Arthur Woolf, all become available, and it is difficult to avoid the conclusion that, observing Hornblower's fate, they were simply waiting for the Watt patent to expire before releasing their inventions.

Now despite the fact that there were many people working in parallel on steam engines, generally without protection of the legal system, and a great deal of overlapping and simultaneous discovery, it is possible that Watt's contribution was so unique and the difficulty of discovery so great that it would not have happened without the promise of a long monopoly. (The facts of the Watt story suggest rather strongly that this was not the case.) But the fact is that Watt would have made a great deal of money even without the legal monopoly. This is strongly indicated by the impact that the expiration of his patents had on Watt's empire. Despite the fact that many new firms sprang up, they produced an inferior engine, and Thompson [1847, p. 110] says that "Boulton and Watt for many years afterwards kept up their price and had increased orders."

In the end, the evidence suggests that Watt's efforts to use the legal system to inhibit competition set back the industrial revolution by a decade or two. The granting of the 1769 and, especially, of the 1775 patents likely delayed the mass adoption of the steam engine: innovation was stifled until his patents expired; and very few steam engines were built during the period of Watt's legal monopoly. From the number of innovations that occurred immediately after the expiration of the patent, it

appears that Watt's competitors simply waited until then before releasing their own innovations in an effort to avoid the fate of Hornblower. Also, we see that Watt's inventive skills were badly allocated: we find him spending as much time engaging in legal action in an effort to establish and preserve a monopoly as he did in actual invention. Our theoretical contention, that innovation may be hurt rather than enhanced through legal monopoly, is given empirical substance through the story of James Watt.

References

Acemoglu, D. and J. Angrist, [2000], "How Large are the Social Returns to Education? Evidence from Compulsory Schooling Laws," in B. S. Bernanke and K. Rogoff (eds.) *NBER Macroeconomic Annual 2000*, pp. 9-59.

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

Aghion, P. and P. Howitt, [1992], "A Model of Growth Through Creative Destruction," *Econometrica*, Vol. 60, pp. 323-351.

Becker, G., [1971], *Economic Theory*, Knopf Publishing Co.

Boldrin, M. and D.K. Levine, [1999], "Perfectly Competitive Innovation," www.dklevine.com, www.econ.umn.edu/~mboldrin .

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

Boldrin, M. and D.K. Levine, [2003], "The Theory of Endogenous Innovation and Growth under Perfect Competition," mimeo, University of Minnesota and UCLA.

Ciccone, A. and G. Peri, [2002], "Identifying Human Capital Externalities: Theory with an Application to U.S. Cities," mimeo, UPF and UCD, December.

Ellison, G. and E. L. Glaeser, [1999], "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?" Harvard Institute of Economic Research Working Paper 1862.

Gallini, N. and S. Scotchmer, [2002], "Intellectual Property: When Is It the Best Incentive System?," in A. Jaffe, J. Lerner, and S. Stern, (eds), *Innovation Policy and the Economy,* Vol. 2, MIT Press.

Gilbert, R. and C. Shapiro, [1990], "Optimal Patent Length and Breadth," *RAND Journal of Economics*, Vol. 21, pp. 106-112.

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

Hart, O. D. (1979), “On Shareholder Unanimity in Large Stock Market Economies,” *Econometrica*, Vol. 47, pp. 1057-83.

Kahn, A. E. [1962], “The Role of Patents,” in J.P. Miller ed. *Competition, Cartels and Their Regulation*, Amsterdam, North Holland.

Krugman, P., [1980], "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, Vol. 70 pp. 950-59.

Leibowitz, S. [1985], “Copying and Indirect Appropriability: Photocopying of Journals,” *Journal of Political Economy*, Vol. 93, pp. 945-957.

Lucas, R. E., Jr., [1988], “On the Mechanics of Economic Development,” *Journal of Monetary Economics*, Vol. 22, pp. 3-42.

Makowski, L. (1980), “Perfect Competition, the Profit Criterion, and the Organization of Economic Activity,” *Journal of Economic Theory*, Vol. 22, pp. 222-242.

Plant, A. [1934], "The Economic Aspects of Copyright in Books," *Economica*, pp. 167-195.

Quah, D. [2002], “24/7 Competitive Innovation,” mimeo, London School of Economics.

Romer, P. M., [1986], “Increasing Returns and Long Run Growth,” *Journal of Political Economy*, Vol. 94, pp. 1002-1037.

Romer, P. M., (1990), “Endogenous Technological Change,” *Journal of Political Economy*, Vol. 98, pp. S71-S102.

Schumpeter, J., [1943], *Capitalism, Socialism and Democracy*, London: Unwin University Books.

Scotchmer, S., [1991], “Standing on the Shoulders of Giants: Cumulative Research and the Patent Law,” *Journal of Economic Perspectives*, Vol. 5, pp. 29-41.

Stigler, G. J., [1956], “Industrial Organization and Economic Progress,” in *The State of the Social Sciences*, edited by Leonard D. White, University of Chicago Press.

Thompson, B., [1847], “Inventions, Improvements, and Practice of Benjamin Thompson, Colliery Engineer, with some interesting particulars relative to Watt's Steam Engine,” reported in J. Lord, *Capital and Power 1750-1800,* London, 1923; www.history.rochester.edu/steam