Should We Give Every Cow Its Calf? Monopoly, Competition and Transaction Costs in the Promotion of Innovation and Creativity

Rufus Pollock
Queens' College
Cambridge University

This dissertation is submitted for the degree of Doctor of Philosophy

January 22, 2008

Abstract

The work presented here is part of a wider research programme oriented around three specific questions. First, how do individual agents appropriate returns from innovation and how is this affected by the availability (or not) of intellectual property rights such as copyrights and patents? Second, how does this translate into the aggregate production of knowledge, once one takes account of the interaction between producers and the cumulative nature of the process of knowledge production? Finally, How can we incorporate this into an estimate of the welfare trade-off inherent in intellectual property rights (the basic prerequisite for formulating rational IP policy)?

The dissertation contains theoretical work on each of these questions together with a brief introductory preamble and a review of the existing literature on the economics of knowledge.

Chapter 3, entitled *Cumulative Innovation, Sampling and the Hold-Up Problem*, examines the interaction of producers through licensing in the context of a cumulative innovation process. With imperfect information about the value of innovations, intellectual property rights can result in hold-up and, therefore, it may be better not to have them. Incorporating 'sampling' by second-stage firms into the basic model, it is shown that, the lower the cost of sampling or the larger the differential between high and low value second-stage innovations, the more likely it is that a regime without intellectual property rights will be preferable.

Chapter 4, entitled *Innovation and Imitation with and without Intellectual Property Rights* presents a simple, yet powerful, model of innovation without intellectual property rights based on first-mover advantage. This chapter introduces a model of imperfect competition in which imitation is costly and demonstrates that, in the absence of intellectual property rights, a significant proportion of innovations may still occur. Furthermore, welfare may be higher in the absence of these rights even though less innovation occurs.

Chapter 5, deals with a more applied problem, namely the form of an optimal copyright regime. Entitled *Forever Minus a Day? Some Theory and Empirics of Optimal Copyright*, this chapter develops several novel general results before turning to the case of optimal copyright term. An estimate of optimal term of 14 years is derived – a level far below that in force in almost all jurisdictions at the present time. Furthermore, this estimate is one of the first which in the literature to be properly grounded in both theory *and* empirics.

Finally, Chapter 6, entitled *The Control of Porting in Two-Sided Markets*, moves away from the economics of knowledge, into the area of (indirect) network effects and two-sided markets. The model it develops, focused on the control of porting by a dominant firm in a platform market, has particular relevance to a variety of past and present antitrust questions.

Acknowledegments

Particular thanks go to my advisors Rupert Gatti and David Newbery who have guided and encouraged me throughout the process of preparing this dissertation. I am also appreciative of the inspiration and time I received from Danny Quah when at a very early stage of this endeavour.
On a more personal note special thanks go to LF for her fanatical proof-reading and (almost) endless willingness to listen. I am also deeply grateful to my parents for their ongoing support and encouragement.

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this thesis have been submitted for any other qualification.

Contents

Chapter 1

Preface

1.1 Background

One of the first printed texts of which we have record is a copy of the Buddhist Diamond sutra produced in China around 868AD. In it can be found the dedication: "for universal free distribution". Clearly, the idea of open access to knowledge has been present since humanity first began to formally transmit and share ideas. It is also likely that the urge to keep ideas secret, particularly those that had 'commercial' value, is equally old.

With the development of trade and technology, particularly during the Renaissance in Europe, these parallel approaches of openness and secrecy continued to evolve but the tension between them also increased. With the introduction of formal monopoly rights such as patents and copyrights during the sixteenth and seventeenth century there emerged a halfway house of sorts whereby the monopoly (and the associated profits) of secrecy was combined with openness in the form of the disclosure of the work.

These alternatives of openness, secrecy and state-sanctioned monopoly have stayed with us down to the present day; while most of our ideas, particularly cultural ones, are 'public domain', free for anyone to use and reuse, a significant portion of the intellectual works and products created by the economies of the world are protected either by some form of intellectual property rights or by secrecy – or by both, as is the case with most proprietary computer software, for example.

However, there have also been considerable changes. On the one hand, there has been a large increase, particularly over the last 30 to 40 years, in the scope and duration of intellectual property rights. On the other hand, and at the same time, especially in recent years, we have seen the rise of self-consciously open models of innovation, particularly in software where the 'copyleft'/open-source approach to knowledge licensing first arose in the 1980s.

However, the most significant of all changes underlies these others, for it is the change in the role of knowledge in society and the economy. Terms such as the 'information age' or the 'knowledge economy' are now commonplace and hard statistics point to the fact that in most Western economies the information-based service sector is now more important than manufacturing. These changes in turn result from, or at least depend upon, a revolution in communication and computer technologies that has greatly reduced the cost of production, distribution and manipulation of knowledge. Whole industries which neither existed nor were imagined 50, possibly even 20, years ago have grown up which exploit these new-found possibilities.

These are vast changes and they have profound implications for the production and dissemination of knowledge, as well as for their regulation and support by government.

My research as an economist is motivated by a desire to understand these developments, and in so doing, to ensure that the policies adopted by governments and others deliver the full benefits of the cultural and industrial creativity available to our society in this digital age.

1.2 Overview of my Research

During the last decade the EU has witnessed a protracted and bitter battle over the issue of software patents. Proponents argued that allowing software to be patented would stimulate innovation whilst opponents (including the majority of software developers) argued it would do the exact opposite.

Being both a software developer and an economist, this issue was of great interest to me. An examination of the literature yielded a confusing picture. On the one hand, the majority of economists working in the area appeared to consider that patents would be damaging in software, due to the nature of the industry – low entry costs combined with a highly cumulative and componentised development model. At the same time, the theoretical (and empirical) literature supporting this view seemed exceedingly sparse. In fact, the vast majority of existing work was dominated by the following assumption and its corollary:

Assumption: Imitation is cheaper than innovation (strong form: imitation is costless – i.e. goods are perfectly nonrival).

Corollary: Innovation is always higher with intellectual property rights (strong form: no innovation occurs without intellectual property rights).

It seemed clear to me, particularly given my own experience as a software developer, that this corollary was wrong – at least for some industries. As such, a central motivation of my research to date has been the question: "Is it possible to find interesting models in which the introduction of intellectual property rights, such as patents, would reduce innovation?" (conversely: "Are there models in which open approaches to knowledge production would be optimal?"). This in turn is part of a wider research programme which can be divided into three parts:

1. How do agents (firms or individuals) appropriate returns from innovation and how is this affected by the availability (or not) of intellectual property rights such as copyrights and patents?

2. How does this translate into aggregate production of knowledge, once one takes account of the interaction between producers and the cumulative nature of the process of knowledge production?

3. How can we incorporate this into an estimate of social welfare that can be empirically determined? Intellectual property rights act to create a monopoly. Whilst this may increase returns to a producer of a piece of knowledge, it simultaneously hinders access (as the price is higher), reducing the benefit to society. Explicitly determining this trade-off, and therefore overall social welfare, is essential to evaluating and formulating policy in this area.

The papers that make up this dissertation include theoretical work on all of these items. Chapter 3, entitled *Cumulative Innovation, Sampling and the Hold-Up Problem*, examines the interaction of producers through licensing in the context of a cumulative innovation process. With imperfect information about the value of innovations, intellectual property rights can result in hold-up and, therefore, it may be better not to have them. Incorporating 'sampling' by second-stage firms into the basic model, it is shown that the lower the cost of sampling, or the larger the differential between high and low value second-stage innovations, the more likely it is that a regime without intellectual property rights will be preferable. Thus, technological change which reduces the cost of encountering and trialling new 'ideas' implies a reduction in the socially optimal level of intellectual property rights.

In Chapter 4, entitled *Innovation and Imitation with and without Intellectual Property Rights*, a simple, yet powerful, model of innovation without intellectual property rights based on first-mover advantage is developed. As stated above, the literature on innovation and intellectual property rights has tended to assume the pure nonrivalry of information goods. Yet, this is at odds with an extensive set of empirical facts, most prominently the evidence that returns from innovation are appropriated primarily via mechanisms other than patents or copyrights and that 'imitation' is itself a costly activity. This paper introduces a model of imperfect competition in which imitation is costly and demonstrates that, in the absence of intellectual property rights, a significant proportion of innovations may still occur. Furthermore, welfare may be higher in the absence of these rights even though less innovation occurs.

Chapter 5 deals with a more applied problem, namely the form of an optimal copyright regime. Entitled *Forever Minus a Day? Some Theory and Empirics of Optimal Copyright*, this chapter develops several novel general results before turning to the case of optimal copyright term. An estimate of optimal term of 14 years is derived – a level far below that in force in almost all jurisdictions at the present time. Furthermore, this estimate is one of the first which in the literature to be properly grounded in both theory *and* empirics.

Finally, Chapter 6, entitled *The Control of Porting in Two-Sided Markets*, presents work on a rather different topic. In contrast to the previous chapters which all deal

with questions related to the economics of knowledge, this chapter builds on another part of IO literature, namely that dealing with (indirect) network effects and two-sided markets. The model it develops, focused on the control of porting by a dominant firm in a platform market, has particular relevance to a variety of past and present antitrust questions ('porting' here denotes the conversion of 'software' developed for one platform, such as Microsoft Windows, to run on another, such as Linux).

Chapter 2

The Economics of Knowledge: A Review of the Literature

2.1 Introduction: The Nature of Knowledge[1]

The starting point for any investigation of the economics of knowledge is the observation that knowledge is different in several crucial aspects from 'normal' physical goods. As emphasized by Arrow (1962), and mentioned by many authors before him,[2] knowledge is:

1. Nonrival (or, at the very least displays significant non-convexities in its production function): in contrast to physical goods, it is, at least approximately[3], costless to reproduce a piece of knowledge once the first 'copy' is made. If one shares a pair of shoes one does not create a new pair – quite the opposite: each party now only has the shoes half the time. However, if one shares a piece of knowledge another gains without any corresponding loss to oneself. As Jefferson eloquently phrased it, over 200 years ago: "He who receives an idea from me, receives instruction himself without lessening mine; as he who lights his taper at mine, receives light without darkening me"[4]

2. Imperfectly excludable (and, in general, greater excludability comes at the cost of more inefficient use – for example in the form of monopoly pricing). The (partial) nonexcludibility could manifest in many forms, as spillovers to other firms, as the inability of a firm to extract more than the monopoly rent from a given product, or even, to take a case emphasised by Arrow, the fact that the seller of a piece of knowledge faces a unique dilemma in that disclosure may be necessary for the sale but may simultaneously eliminate all demand.[5]

Together these lend knowledge the aspect of a 'public good': from the viewpoint of society, once a piece of knowledge is in existence the optimal thing to do is distribute

[1]There have been long-running debates about the distinction between invention and innovation, and between technology and science – as well as whether such distinctions serve any valuable purpose. It is not my object to engage in these discussions here. Instead, I shall assume all innovation is related to the creation of new 'knowledge' – including the knowledge of how to develop associated applications (in this we follow the lead of Foray (2004)).

[2]From the academic literature an early example is the work of Plant (Plant, 1934a,b). As evidenced by the Jefferson quote below, as well as the widespread presence of early modern forms of intellectual property, there has clearly been some awareness of the special nature of knowledge from the very earliest times. However initial thinking on the subject, even among economists was hampered by a lack of clear understanding of the nonrivalrous and nonexcludable nature of knowledge – as well as the interplay between the two (see pp. 17ff of Hadfield (1992)).

[3]We shall return to how accurate this approximation is in some detail below. See, for example, Sections 2.6 and 2.6.2.

[4]Jefferson to Isaac McPherson 13 Aug 1813, Jefferson (1905) vol. 13 pp.333–335.

[5]Knowledge, by its nature, tends not to be a homogeneous good and thus a buyer is unlikely to be willing to pay much for a piece of knowledge whose properties are unknown. However, if the seller reveals the knowledge to the buyer in order to demonstrate its value the very act of disclosure serves as to transfer the knowledge and eliminate the seller's market.

it at marginal cost (which may be zero or very close to zero). At the same time the extreme heterogeneity (and uncertainty) associated with knowledge, as well as its close connection to the production and development of other goods make it hard to adopt a pure 'central-planner' approach of up-front funding (based on taxation) followed by free distribution which is the method adopted for other public goods such as defence – though, of course, much knowledge production is funded in this manner (including this very paper).

The central point to take from this is that due to the special nature of knowledge, its production and distribution can *not* be optimally organized via the free workings of a decentralized market system. Consequently, this is an area of economic analysis which necessarily has a particularly close relation to questions of regulation and policy – be they the optimal form for the intellectual property system to take, or the level of public expenditure on R&D. As Arrow summarized, writing 30 years after his original paper (Arrow, 1993): "knowledge is a hard commodity to appropriate, and it is socially inefficient to appropriate it." This dilemma continues to haunt economists and policymakers today.

2.2 Early Work and Patent Race Models

Following Arrow, there was scattered early theoretical work looking at various aspects of the 'R&D question', but the absence of game-theoretic tools and the breadth of the field meant that progress was limited and showed little consistency in approach.[6]

However, beginning with a series of papers in the late 1970s and early 1980s (Loury, 1979; Lee and Wilde, 1980; Dasgupta and Stiglitz, 1980a; Reinganum, 1981), there was sustained attention in the industrial organization literature to developing new 'micro-founded' models of R&D using non-cooperative game-theory techniques. A long line of work, particularly focused on 'patent race' style models, arose whose basic features

[6]Examples of earlier literature include Horowitz (1963), who investigates the incentives for R&D by a single firm in a n-player Cournot model; Scherer (1967), who examines R&D rivalry in a Cournot-style duopoly; Barzel (1968) who raises the possibility that competition induces too early introduction of new technology; (Kamien and Schwartz, 1972a) which investigates in a decision-theoretic framework the optimal R&D program for a firm as a function of market conditions (degree of rivalry etc); Kamien and Schwartz (1972b) which looks at the impact of the degree of rivalry (and imitation) on the innovation level of a 'leader' firm; Kamien and Schwartz (1974), which examines the impact of rivalry on optimal patent design; Kamien and Schwartz (1978) which again looks at the effect of rivalry on the optimal innovation strategy of a given firm (many of these papers by Kamien and Schwartz are collected along with additional material in (Kamien and Schwartz, 1982)).

The main feature of this work compared to that which came later was the absence of any modelling of the strategic interaction between firms. Thus, for example, the long series of papers by Kamien and Schwartz focus only on the optimal behaviour of a single firm as a function of an exogenously given environment – even when rivalry by other firms is an explicit consideration.

were:[7]

1. A focus on the supply-side. These were models of R&D races and the focus was on the suppliers of R&D. The demand-side both in the form of end consumer demand and other firms (licensing) were often black-boxed – in general one would simply assume that a given R&D project would yield income v with social welfare being $w \geq v$.[8]

2. A known R&D goal (the discovery) shared by all the participants in the race.[9]

3. A known functional form linking expenditure with discovery.[10] Where the race was dynamic this would entail the use of a memoryless (poisson-style) discovery function (that is the probability of making the discovery only depends on current expenditure and not on past expenditure).[11]

4. Rational, strategic behaviour on the part of the firms engaged in R&D and innovation.[12]

Even with these simplifications (which were most significant in the area of the demand-side structure and product market competition), there are still multiple factors which generate a divergence between social and private incentives and provide areas for investigation. For example, the difference between social and private returns (problematic because of the combination of imperfect appropriability *and* non-convexities), the winner-takes-all aspect of knowledge discovery which results in 'pool' externalities, the divergence between social and private attitudes to risk, the presence of uncertainty and asymmetries in information which give rise to a host of moral hazard and adverse selection problems.

[7]Of course like any generalisation this will not be entirely accurate and we will discuss some of the places where papers differ from this below.

[8]Though e.g. Dasgupta and Stiglitz (1980a,b) both consider explicit models of the product market – perhaps assisted by the fact that these are process innovation models. In particular, Dasgupta and Stiglitz (1980b) develops a fairly detailed two-stage model in which, after choosing cost-reducing R&D expenditure, firms play a standard Cournot game.

[9]That is all firms were attempting to develop the same product or to develop a process innovation for the same product. As such there was little incorporation of the possibility of product differentiation whether of a horizontal or vertical kind (as demonstrated, though not in specific R&D context by Gabszewicz and Thisse (1980); Shaked and Sutton (1983)). However, see Beath, Katsoulacos, and Ulph (1987) for an exception to this tendency.

[10]Scotchmer (2004) for thoughtful discussion of the advantages and disadvantages of 'production-function' style models.

[11]Some work was done on dynamic models that did not assume a memoryless discovery function, for example Fudenberg et al. (1983) and Grossman and Shapiro (1986) as well as the series of papers by Harris and Vickers e.g. Harris and Vickers (1985a,b, 1987). Reinganum (1982) models a dynamic race but assumes a fixed end-point ('doomsday') by which innovation must occur.

[12]An explicit alternative to this approach can be found in the 'non-optimizing', evolutionary, models developed by Nelson and Winter (1982) and subsequent authors.

Starting from these considerations a large body of work investigated a variety of questions the most important of which went back to Schumpeter (1947): "What is the market structure which maximises innovation", or alternatively: "Is competition conducive to technical advance?". Specifically:

2.2.1 How does the amount of R&D per firm vary with the number of firms?

It decreases according to Loury (1979); Dasgupta and Stiglitz (1980a,b); Delbono and Denicolo (1991);[13] but increases according to Lee and Wilde (1980) (using a slightly modified version of Loury's model) and Reinganum (1982, Prop 6.)[14]; and remains unchanged according to Sah and Stiglitz (1987), who allow firms to choose the number of R&D projects as well as the effort per project (this resembles the situation of a monopolist in many of the other papers).[15]

2.2.2 How does total amount of R&D vary with the number of firms?

It increases according to Loury (1979) and Lee and Wilde (1980) but may decrease or increase depending on other factors (such as barriers to entry) according to Dasgupta and Stiglitz (1980b), or even be invariant to market structure according to Sah and Stiglitz (1987).[16]

2.2.3 How does this level of R&D relate to optimum.

Are R&D programmes chosen by competitive firms too risky or not risky enough? Not risky enough according to Dasgupta and Stiglitz (1980a), and Judd (1985) (who generalises to a dynamic GE type framework);[17] too risky according to Dasgupta and Maskin (1987, Prop. 5); and either too risky or not risky enough (depending on the

[13]Delbono and Denicolo (1991) present a model very similar to Dasgupta and Stiglitz (1980b) (who, surprisingly, they do not cite) in which there is process innovation with Cournot competition. As a result more firms have two countervailing effects: under Cournot competition more firms mean lower payoffs to the winner of the R&D race which damps effort, but at the same time one still has the 'pool' externality which drives up effort.

[14]Though as she points out in the following commentary by varying the rewards of imitators versus innovators it is easy to construct examples that go the other way.

[15]Though this invariance result is shown to depend strongly on Sah and Stiglitz's choice of Bertrand competition in the product market – see Farrell, Gilbert, and Katz (2002).

[16]Though see previous footnote.

[17]Later work, such as Cabral (1994), also obtains similar results.

skewness of the distribution of returns) according to Bhattacharya and Mookherjee (1986).

Are R&D programmes too correlated? Yes, according to Dasgupta and Maskin (1987). No according to Bhattacharya and Mookherjee (1986).[18] What are optimal subsidies or taxes? Dixit (1988) considers this in detail.

2.2.4 How do incentives vary with market structure?

Specifically, what are the relative incentives to undertake R&D of (a) an incumbent monopolist (b) an incumbent monopolist facing new entrants (c) new entrants with no incumbent? Consideration of (a) and (c) was the focus of Arrow (1962), who concluded that due to a 'replacement' effect an incumbent monopolist had less incentives than entrants to do R&D. Dasgupta and Stiglitz (1980a) provide a detailed examination of all of these possibilities and a consideration of (b) versus (c) form part of the next item.

2.2.5 Does monopoly persist?

Specifically do incumbents retain their position (persistence of monopoly) or are they 'leap-frogged' by new entrants in a form of Schumpeter's creative destruction? Additionally, does an incumbent (leader) have more or less incentive than an entrant (follower) to spend on R&D?

Gilbert and Newbery (1982) investigate the incentives of a monopolist to engage in 'pre-emptive' patenting (and associated patent shelving) in order to preclude entry. The basic idea[19] is a very simple one: the rents accruing to a monopolist are always at least as large as the total rents available under any other market structure (including duopoly). Thus, all other things being, equal the incentive of a monopolist to remain a monopolist is at least as large than for an entrant to become a duopolist. However, as other authors pointed out, all other things need not be equal. For example, Reinganum (1983), examined a similar situation but in the patent-race framework. In that situation higher spending by the monopolist serves not only to increase the probability that she wins the race (good for the monopolist), but also to hasten the point at which the race ends (bad because she is the current incumbent). As a result, despite the higher rents available to the monopolist it is possible that the monopolist will spend less than entrants and hence will be less likely to win the race.[20]

[18] A recent paper, Bulut and Moschini (2006), has shown how the availability of multiple instruments, for example trade secrets in addition to patents, may ameliorate the problem of excessive correlation.

[19] Very similar to that presented in Dasgupta and Stiglitz (1980a, p. 13).

[20] On a different tack, Salant (1984), in a comment on the Gilbert and Newbery paper, points out that under efficient ('Coasian') bargaining all patents (old and new) will always end up under the control of this firm (precisely because such an outcome maximises rents) and that, as a result, it is

Turning to the multi-stage case with certainty in R&D, Fudenberg et al. (1983) and Harris and Vickers (1985a,b), both establish an even more extreme version of this result, in which, whoever has the advantage in the race – be it in terms of valuing the prize more, being better at R&D, or having made greater progress so far, completely dominates (for example, the 'follower' may simply drop out with the 'leader' behaving as if there were no competition).[21] This result arises from the combination of subgame perfection and certainty. Relaxing these assumptions in the second part of their paper, Fudenberg et al. (1983) investigate what occurs when the participants in the race do not know immediately the level of their competitor's effort. They show that this increases the level of R&D both on the part of the 'leader' and the 'follower', furthermore it may allow 'leap-frogging' – a situation in which the 'follower' jumps from behind to the take the lead.

Extending their earlier results to the case of uncertainty in the R&D function, Harris and Vickers (1987) establish that, in general, in a single-stage and multi-stage race (Property 3.1 and Property 4.2) the 'leader' expends more effort than a 'follower'. They also find that in both cases total effort increases as the deficit between the two competitors narrows.

In contrast to this line of papers which implied that incumbency would persist (dominance would increase), Reinganum (1985), which explores a multi-stage patent race model, finds that an incumbent monopolist would spend less than its rivals and that, as a result, there would be a pattern of repeated monopoly with each monopolist being displaced in turn by a new entrant.[22] In a similar vein, Vickers (1986),[23] shows that one can obtain either increasing dominance (one firm extends its technological superiority) or have 'action-reaction' (technological leadership repeatedly changes hands). In his model, which features certainty in R&D, which outcome obtains depends on the form of product market competition: 'tough', Bertand-like, leads to increasing dominance while 'soft', Cournot-like, leads to 'action-reaction'.[24]

---

not *necessarily* the case that the monopolist has higher incentives to do R&D than an entrant.

[21]Some of the subtleties of the analysis are necessarily lost in a summary such as this. Specifically, Harris and Vickers introduce the concept of 'safety' and 'trigger' zones. In a 'safety' zone a player behaves as if in the absence of competition (and the other player bids 0) while in a 'trigger' zone, whoever's turn it is to move must win immediately (to prevent the other player winning).

[22]Given the different conclusions reached by these different sets of models it is perhaps interesting to note one of the few empirical studies of the topic, that of Lerner (1997), finds that, in the area he studies (the computer disk-drive industry), firms that trail the leader appear to innovate more.

[23]Extended to the case of incremental innovations by Delbono (1989).

[24]The question of whether a dynamic oligopoly will produce 'increasing dominance' or 'catch-up' is a general one, not specific to the area of R&D. See, for example, the general model and results of Budd, Harris, and Vickers (1993).

2.2.6 Summary

As is often the case, it turned out the answer to many of the questions posed was: 'it depends on the model' (or the particular values of parameters in the model).[25] Nevertheless, despite variation in results on many of these issues there was substantial consistency on some basic things. Most importantly, R&D levels could be both too high as well as too low.[26] In particular, competition *in R&D* in a winner-takes-all framework generates increased (even excessive) incentives compared to a monopoly situation.

2.3 Patent 'Design'

The R&D literature discussed above tended to ignore exactly how the innovator's rents were obtained and how those rents depended on the intellectual property (patent or copyright) regime. After all, there were plenty of other factors to examine.

Nevertheless, the socially optimal design for patents, particularly the examination of the trade-off between the benefits of increased innovative activity and the costs in the form of deadweight losses, had been present from the earliest point[27] and it was soon after the appearance of Arrow's paper that Nordhaus (1969) provided the first attempt formal model of optimal intellectual property policy. Nordhaus, (reinterpreted and expanded upon by (Scherer, 1972)), looked at the basic case of a process innovation that reduced the marginal cost of a production process and sought to determine the optimal patent term – the point at which the marginal benefit of increased protection in the form of incentives for a firm to invest to develop a cost saving innovation would exactly equal the extra deadweight losses to society of granting that firm monopoly power for longer. Nordhaus showed, that while (obviously) determining any explicit

[25]Furthermore, clearly any specific results on items such as taxes and subsidies (e.g. Dixit (1988)) would depend on knowledge of a full array of private as well as public information. In such a situation one must ask why, if the social planner has access to such detailed information, he or she does not simply up-front fund the research and avoid the inevitable dead-weight losses associated with private, patent or secrecy-based, R&D.

[26]The basic reason for this had long been known: on the one hand, firms do not extract the full surplus they generate, while on the other hand, competition encourages the premature introduction of innovations. See for example, Barzel (1968) who noted (p. 348): "It is widely recognized that when innovators are unable to realize the full benefits generated by their innovations the profit motive may not provide an incentive strong for them to innovate at the socially optimal rate. On the other hand, it has *not* been recognized that competition between potential innovators to obtain priority rights (and profits) from innovations can result in premature applications of discoveries. [italics added]"

However, as emphasized by Dasgupta (Dasgupta, 1988), putting this result on a sound theoretical footing was one of the major achievements of the first wave of game-theory, 'patent-race' style, models – along with a include a clearer understanding of the non-additive nature of parallel research (see Dasgupta (1989)).

[27]For example, Plant (1934a,b) clearly considers this, as does Hurt and Schuchman (1966).

value for optimal term depended on the parametrisation of the model, one could at least show that optimal term would be finite.[28]

Following the development of the substantial literature on patent races discussed above, the focus widened in several directions. First, there was an examination of the potential for competition in the end product market via imitation. This tied back very directly into the optimal patent literature started by Nordhaus with the added complexity of considering patent breadth as well as patent length. Second, and relatedly there was the question of licensing – compulsory or voluntary.[29]

Once competition in the end product of R&D is to be permitted, due to imitation for example, one needs a model for this competition and how it impacts on the innovator's income (usually one assumes the innovator obtains a patent and so the question is then how its strength vis-a-vis imitators – the breadth of the patent – impacts on the flow of rents per unit period). There are two basic approaches, which we might label the 'reduced form' and the 'microfoundations' (location-model) respectively.

As the names indicates the first approach, taken in e.g. Gilbert and Shapiro (1990); Gallini (1992), involves 'black-boxing' the impact of patent breadth on the patentee's rents in a single functional form, while the second approach, used e.g. in Klemperer (1990); Waterson (1990), involves the provision of an explicit model of the relation between imitators and innovators – usually based on a locational model of some kind (if one is to allow competition in the product space but still retain the concept of excludability one needs a product space which is at least 1-dimensional). These various approaches and assumptions yield a fairly diverse set of results, which are not always consistent.

Gilbert and Shapiro (1990), under an assumption that patent breadth is increasingly costly in welfare terms, find that patents with finite width but infinite length will be optimal.[30] Klemperer (1990) by contrast (as Gilbert and Shapiro themselves note), has a more complex situation in which, at least in some circumstances, optimal patents

[28]As emphasized by Horowitz and Lai (1996) this model also implied that the innovation-maximizing patent length exceeded the welfare-maximizing patent length (a point made in relation to copyright by (Landes and Posner, 1989)). Horowitz and Lai study a more general case with multiple patent races and where firms choose both effort and size of innovation. They find that market leader innovates more near patent expiry but that extent innovation is an increasing function of patent term. Overall they establish a similar result to Nordhaus, showing that patent length should be finite and shorter than the level that maximizes the level of innovation. An obvious point perhaps but one often ignored by policymakers.

[29]We shall return to this subject in greater detail in Section 2.5.

[30]This has some analogies with the earlier paper of Tandon (1982), who investigates the interaction of compulsory licensing and patent length (just like patent breadth compulsory licensing limits the price a patent holder simultaneously increasing rents and deadweight losses). Similarly to Gilbert and Shapiro (1990), Tandon finds that optimal policy involves patents which are infinitely lived but whose price is limited via the price of the compulsory license.

are broad but short. In Klemperer's model all demand is situated at a single point in product space at the same location as the patented product. Breadth is then naturally interpreted as an exclusion radius. Free entry (with zero costs) imply that the patent-holder's price will be then be limited to the cost of 'transport' to the competitive fringe firms on this radius. Thus, in Klemperer's model welfare losses arise not only from the pricing decision but also from travel costs incurred by consumers and the design of the optimal patent must trade-off these two losses.[31] As a result, depending on one's assumption about the distribution of valuations and transport costs, one can have an optimal patent being either long and narrow (Prop. 1 and 2) or, short and wide (Prop. 3 and 4). In the general case the trade-off between width and length will be determined by the relative elasticities of the distribution of values and transport costs.

In a similar vein, Waterson (1990) uses a simple Hotelling line model of product space and considers how the breadth of the patent, interpreted as a simple exclusion zone, impacts (via litigation) on the imitator's behaviour and, thereby, on the innovator's profits and (socially beneficial) product differentiation. Waterson finds that the optimal regime depends on the importance of variety: where it is important narrower patents are desirable, as they allow for entry and thereby increase the number of products on offer, but when variety is less important broad patents, which prevent imitation, are best, as they maximize innovator's rents. Here, in contrast to Klemperer, but in line with standard locational models, all consumers purchase – be it from the innovator or the entrant/imitator, and so the welfare effects arise solely from transport costs and the incentive provided to the innovator.

Meanwhile, Gallini (1992) emphasises strategic considerations of imitators. If patent length is made longer while breadth is reduced this increases the incentives of imitators to 'invent-around' the patent and thus reduces the patent's actual (as opposed to statutory) life. As such, increasing length and reducing breadth may not, in fact, be optimal. Rather, broadening patent protection but shortening its term, by reducing the incentives of imitators/innovators to invent-around, will provide the optimal way to deliver rents to the patent-holder at least cost to society.[32]

Finally, Denicolo (1996) introduces the possibility that many firms race for a patent and shows how the variety of existing results primarily stem from differing assumptions about the structure of the product market. In particular, he provides simple models in which all combinations of maximum breadth and minimum length, mini-

[31]Optimally one would want all consumers to purchase from the patentee and thereby incur zero travel costs but with unobserved heterogeneity in valuation and/or travel costs the monopolist may set a price above the outside option of some consumers.

[32]Note, that 'inventing-around' will result in just the kind of wasteful duplication of effort that underlie the 'pool' externalities of the standard patent race literature.

mum breadth and maximum length and neither maximum nor minimum breadth of length are optimal.

2.3.1 Patent Menus

The 'patent design' literature discussed so far has focused on picking a single optimal value for one or several of the patent parameters. An obvious extension to this approach as been to allow a *menu* of possible values for, say, patent length combined with a set of associated fees – in fact most patent offices already had a system like this in which fees were charge annually for the renewal of the patent (though their motivation for this approach rested on simple cost sharing rather than any consideration of mechanism design).[33]

Thus, for example, Scotchmer (1999); Cornelli and Schankerman (1999), examine the case where a menu of lengths and fees are offered and show that a menu can be welfare improving. Cornelli and Schankerman (1999) show, furthermore, that a renewal system (as opposed to a simple fee) offers additional advantages if there is ex ante uncertainty about the value of a patent. Hopenhayn and Mitchell (2001) extend the menu approach to the case where firms can also choose the breadth of their patent and show that by trading off breadth and length one may not require fees at all (though, compared to length, it is harder to see how, in practice, a policy-maker is to offer a variety of breadths).

2.3.2 Novelty and Non-Obviousness

There are usually considered to be three requirements for a discovery to be patentable: it must not fall within an excluded subject area (for example, up until very recently most jurisdictions excluded business methods from patentability); it must be novel; and it must be non-obvious (have an inventive step).[34] Thus various authors have looked at varying some of these other characteristics, in particular novelty and the non-obviousness requirement (the size of the inventive step). From an economist's perspective, particularly when developing a model, it is not clear that novelty adds anything beyond non-obviousness – anything which is non-obvious represents an advance beyond what is currently known and therefore must, a fortiori, be novel.[35] Furthermore,

[33]Note also, that for some time before this theoretical attention several empirical papers, most notably Pakes (1986); Schankerman and Pakes (1986); Lanjouw (1998); Schankerman (1998), had used patent renewal data to estimate the distribution of patent values.

[34]Some jurisdictions, such as the US, also include a requirement – rather similar to the first – that the invention must be susceptible of industrial application.

[35]Even an article such as Scotchmer and Green (1990) which contains 'Novelty' in its title is, in actual fact, an article about the size of the inventive step.

novelty is a purely binary concept: either an invention is novel or it is not and it would seem odd indeed to have a patent system which allowed the (re-)patenting of existing work. Thus, our focus reduces to that of considering the non-obviousness/inventive step requirement. Given that the concept of an inventive step implies some form of advancing (cumulative) line of innovation, we shall defer the question of its optimal design to the next section, which deals exclusively with that subject.

2.4 Cumulative Innovation

The idea of multi-stage patent races was present in some of the early work[36] but the focus was on the differing behaviour of participants over the entire race. Furthermore most of these approaches, at least implicitly, assumed technological independence between stages (and between firms) and focused instead on strategic dependence between stages (the same firms took part in the different stages). For example, as discussed above, a long line of models considered how an incumbent monopolist competed against new entrants to develop a new drastic (or non-drastic) process innovation. Even in those models which explicitly incorporated multiple stages there was little sense that a new innovation 'built-upon' the old – there was, for example, no requirement that a new innovation be sufficiently 'big' in order to qualify for protection, or for a new innovator to obtain a license from the owner of the previous innovation.

By contrast, the cumulative innovation literature discussed here emphasized the technological dependency between stages – while allowing greater strategic independence (the set of firms participating in different stages were often completely unrelated). In a manner similar to the quality-ladder literature innovations were considered as advancing along some set of quality dimensions (usually one for simplicity). Patent breadth could then be (re-)interpreted[37] as distance along this line, and, depending upon the structure and strength of intellectual property rights, new innovators might infringe existing rights and therefore require a license to produce. Furthermore, new externalities (or old ones in new forms) arise: are early innovators, who develop the base upon which future developments will build, adequately compensated for the potential they create (its 'option' value);[38] will existing innovators inefficiently exclude

[36]Many of the models considered cases in which new entrants competed with an incumbent to develop a new innovation (see above section for references), and many models, e.g. Reinganum (1985); Harris and Vickers (1987); Vickers (1986); Delbono (1989), had explicitly modelled multi-stage innovation.

[37]One could see the previous literature on patent breadth, discussed above, as focused on horizontal differentiation (being different), while cumulative innovation dealt with vertical differentiation (being better). Of course, the distinction between better and different in many cases is a fine one.

[38]This question clearly has important analogies with the literature on general spillovers, see Spence (1984), D'Aspremont and Jacquemin (1988) etc.

those who might extend their work?

It is in this sense that cumulative innovation 'ups the stakes': in traditional models of optimal patenting the only limit on infinite patent length was monopoly cost. These costs are generally thought to be relatively small.[39] Given this, and that private firms – even with intellectual property rights – are unlikely to extract anything close to the total social surplus of the innovations they produce, this would imply that patents should be as long as possible (and fairly wide too). But once we have cumulative innovation we need to consider the impact of higher 'prices' (in the form of licensing and transaction costs) on future innovators. Furthermore, when a given innovation does not occur society loses the entire social surplus – which may be very large.[40] Thus, with cumulative innovation, because intellectual property may result in lost follow-on innovation as well as traditional deadweight losses, the costs of too much intellectual property may be substantially higher than in a single innovation context. Conversely, the costs of too little intellectual property may also be much larger: when a given (first-stage) innovation fails to be made because expected rents are too low, society loses not just the value of that innovation but the value of all innovations that would have built upon it.

In terms of the literature, a general awareness of cumulativeness of knowledge has been present from a very early point – it is after all, an omnipresent phenomenon in most areas of human enquiry. However there was little formal modelling prior to the early 1990s.[41]

Initial surveys include Scotchmer (1991), who focused on the first issue mentioned above (that early innovators may not be adequately compensated for the 'option-value' of their innovation), and Merges and Nelson (1990) who provide a multitude of examples that demonstrate in relation to the second issue ('hold-up' or exclusion of follow-on innovators). However the first formal models was provided by the paper of Chang (1995) and Green and Scotchmer (1995). Chang examines the situation where a new innovation builds upon an old and the stand-alone values of the two innovations differs (the stand-alone value is that which would obtain in the absence of infringement and associated licensing). He shows that optimal breadth[42] is not a monotonic function of

[39]Initial work on this by Harberger (1954) (extended by Schwartzman (1960)), found relatively small welfare losses from monopoly. That said, recent work on the effect of patents in the Indian pharmaceutical industry (Chaudhuri, Goldberg, and Jia, 2006), found very large effects, with consumer welfare losses over ten times producer gains (though note that in their paper losses arise not just from higher prices but also from a loss of choice).

[40]An analogous point is made in the context of trade barriers by Romer (1994).

[41]There are of course exceptions: for example the 'prospect theory' of Kitch (1977) implicitly assumes some form of cumulative innovations (otherwise there is nothing for the initial patent-holder to coordinate).

[42]Breadth here is interpreted in terms of the probability that the second-stage product infringes on

the relative values of the two-products. Courts should grant broad protection both to first-stage products that are very valuable relative to the improvements and to those that have very little (stand-alone) value relative to improvements.

Green and Scotchmer, focus on the possibility of ex ante licensing, and how the interrelation of ex ante vs. ex post licensing payoffs affect the incentives of the innovators at the two stages. In their model, in contrast to Chang, first-stage firms have perfect information about second-stage firms' values and costs. This combined with efficient bargaining eliminate any possibility of 'hold-up' (the situation where second-stage firms do not invest either because of anticipated or actual hold-up of their investment at the licensing stage). As a result Green and Scotchmer tend to find that very broad (even infinite) protection is optimal.[43]

Building on this framework Scotchmer (1996) points out that if second-stage innovations are patentable, even with multiple competing second-stage firms, a first-stage innovator will be unable to extract the full 'option-value' of the innovation. Bearing this in mind, she asks whether in some cases it might not make sense to increase the bargaining power of first-stage firms even further by making second-stage innovations unpatentable (in which case, under perfect information, a first-stage firm can extract all surplus).[44]

Denicolo (2000) extends the model of Green and Scotchmer to incorporate patent races at the two innovation stages (though he simultaneously simplifies the licensing aspect). Here stronger intellectual property (greater breadth) in the form of second-stage products being infringing increase the rate of first-stage innovation but retards the rate of second-stage innovation (compared to the non-infringing case). Welfare (and policy considerations) are complicated by the fact that in addition to this trade-off one must also incorporate the effect of 'excessive' incentives generated by racing as well as the opposing effect generated by the gap between the private and social value of innovations. Denicolo finds that it will generally be better to make second-stage products patentable (in contrast to (Scotchmer, 1996)) and that in some cases breadth should be wide (second-stage products infringe) but in others the breadth should be narrow (second-stage products do not infringe).

Matutes, Regibeau, and Rockett (1996), introduce yet another, subtly different,

the first firm's patent and vice-versa. Thus a large breadth means the second-stage product is likely to infringe while a narrow breadth means it is likely not to.

[43]Though, interestingly and rather counter-intuitively, with uncertainty (but symmetric information) about second-stage costs a shorter (finite) breadth may be optimal. The reason being that a reduction in breadth increases the ex post payoff of the second-stage firm. In so doing it reduces the ability of the second-stage firm to threaten to not invest (because of the hold-up risk) and therefore shifts bargaining from the ex ante to the ex post stage. This benefits first-stage firms and the overall result is to increase innovation.

[44]Both in assumptions and conclusions this result has some analogies with Kitch (1977).

version of breadth. In their model there is a single 'basic' innovation of little value by itself but with many valuable applications (furthermore the creator of this basic innovation may conceal its existence for some period while developing applications). 'Breadth' is then the number of these applications that are 'reserved' for the owner of the basic patent. They then contrast the breadth of protection (how many applications are 'reserved' in perpetuity) versus the length of protection (protect all applications for some fixed period T). They find a general preference for the breadth rather than length approach, primarily because of its effect on eliciting early disclosure.[45]

Meanwhile O'Donoghue, Scotchmer, and Thisse (1998) present a very rich model which incorporates: (a) patent-race style (poisson-process) arrival of ideas in a cumulative chain (so each idea represents an improvement on the current innovation); (b) a distribution of idea values (so not all ideas become innovations); and (c) a homogeneous or heterogeneous set of consumers.[46] By having an infinite sequence of ideas (and related innovations) the authors are able to differentiate between leading breadth (how new a new innovation must be to be non-infringing) and lagging breadth (how far behind the current state of the art an imitator must be to be non-infringing). They consider various possible policies, specifically full lagging breadth only, infinite leading breadth plus finite patent life, finite leading breadth plus infinite patent life (note that in this context a patent's life may terminate either when the patent expires or when it is superseded by a new non-infringing innovation and thus both these cases correspond to finite *effective* patent life). They show that in the simple (non-oligopoly) model lagging breadth alone will not provide sufficient incentives for R&D[47] and that either of the alternative policies may provide a remedy (though with subtly different effects on welfare).

Returning to a simpler Green and Scotchmer style model, Bessen (2004), focuses on ex ante licensing combined with asymmetric information about the values/costs of second-stage innovations (which are not known to first-stage firms). As a result hold-up can occur: first-stage firms will set the royalty rate to maximize expected royalty income and this rate will be above the level some second-stage firms are willing to pay. As a result there will be a trade-off between transferring rents to first-stage firms (which encourages innovation at that stage) and the hold-up of second-stage firms

[45]Referring back to the basic 'breadth' literature discussed in section 2.3 this can be seen as similar to 'finite-breadth/infinite-length' type regime recommended by e.g. Gilbert and Shapiro (1990).

[46]This second case permits oligopolistic competition in a vertically differentiated market a la Gabszewicz and Thisse (1980) and Shaked and Sutton (1983).

[47]Though, perhaps rather surprisingly, in the richer oligopoly model with heterogeneous consumers these need not be the case. Here because firms enjoys rents both as leaders and followers, lagging breadth alone may be sufficient to elicit efficient investment if new 'ideas' are 'infrequent' (see Prop. 5 p. 18).

(which reduces the level of second-stage innovation). Looking at the optimal policy in the form of an exogenously (society-determined) ex-post royalty rate, Bessen shows that the optimal level of such a royalty is below that chosen ex-ante by first-stage firms. As a result in his model all licensing occurs ex post at the societally-determined level – a finding he interprets as fitting with the empirical work of Anand and Khanna (2000) on the structure of licensing contracts.

In recent, as yet unpublished work (Bessen and Maskin, 2006), Bessen along with co-author Eric Maskin, extends this model to case where there is a sequence of innovations. At each stage there are (the same) two firms, each of which may choose to participate or not in researching the current innovation. The next innovation stage is reached if, and only if, research at the current stage is successful and success is an increasing function of the number of participating firms and the authors consider two possible regimes: one in which there are patents and one in which there are not. With patents the patent-holder can extract the full value of the innovation and, because subsequent innovation are assumed to infringe, allow the patent-holder to extract a license fee from the follow-on innovator. Without patents both the winner and a loser of a given stage receive a fraction s of the innovation value and the winner has no rights over subsequent stages. Finally, and importantly, just as in the original model there is asymmetric information about costs: firms come in two cost-types and the cost-type is only known to the firm and not its competitor.

As a result of this asymmetry of information, when patents exist, a patent-holder may set a royalty-rate which is too high for a high-cost firm to participate. As a result that firm will be excluded from participation in future innovation stages and the value of this participation is thus lost. As a result there are costs as well as benefits to having patents and as Bessen and Maskin show (Proposition 7, p. 29) in some circumstances (a sufficiently dense tail to the distribution of innovation values and a low enough probability of a low cost innovation) the costs may outweigh the benefits and a regime without patents will yield more innovation (and social value).

2.4.1 Inventive Step

As already discussed, cumulative innovation models were often used to evaluate policy, particularly in relation to the vertical breadth of protection. Another natural, but more specific, application was in the evaluation of the inventive step requirement (see Section 2.3.2 above). An early paper was that of Scotchmer and Green (1990) which though formulated in terms of novelty was in essence about the size of an inventive step. Contrasting a strong with a weak novelty requirement the paper mainly focused on the strategic impact of disclosure on discouraging firms from patenting small improvements

even when this was possible under a weak novelty scenario.

Looking more directly at the inventive step issue, van Dijk (1996) investigates a duopoly model of vertical (quality) product differentiation in which an 'imitative' firm can choose the size of its improvement to the original innovator's product and the choice is constrained by the size of the inventive step. Van Dijk shows that a low inventive step makes no difference to the choice of an imitator, a medium inventive step actually harms the innovator by 'committing' the imitator to a higher level of effort and a high inventive step benefits the innovator by blockading the market completely and leaving them in a monopoly position.

Turning to an infinite sequence of innovations, O'Donoghue (1998), develops a 'quality-ladder' model with an infinite sequence of patent races in which firms may choose both their effort and the size of the targeted innovation.[48]. The technological leader alone makes a profit and this profit is a function of the difference between the quality of her innovation and the next best available. In addition to leading and lagging breadth O'Donoghue considers the size of the inventive step and shows that a patentability requirement (a minimum inventive step size) can stimulate innovation because it extends the effective life.[49]

Hunt (2004), develops a similar model though he endogenizes entry (using a fixed entry cost) and makes the size of a given innovation exogenous. Hunt's central result is that the rate of innovation is an non-monotonic function of the inventive step with a unique inventive step size that maximizes the rate of innovation. This is due to the interaction of two competing forces: on the one hand a larger inventive step makes it more likely that a firm's research efforts will yield no profits (because the invention will not be patentable) but on the other hand it extends the period of incumbency for a firm that does obtain a patentable invention. For similar reasons, Hunt also finds that in his model an industry with faster technological progress should have a higher inventive step.

2.5 Licensing

The question of licensing is an important one – and of much more general concern than simply in its relation to cumulative innovation. Questions that arise include why and when firms will license, the structure of licensing contracts, and the effect of licensing on

[48]This approach is very similar to O'Donoghue, Scotchmer, and Thisse (1998) (see above) and builds upon the approach developed in the endogenous growth literature by Aghion and Howitt (1992); Grossman and Helpman (1991a).

[49]Though there are some subtleties: for example, in the case of a nonlinear profit function a minimum inventive step, while increasing the efforts of followers, may have an ambiguous effect on the leader's incentives (see Proposition 4, p. 670).

R&D incentives and welfare. Prominent examples of work on these topics is provided by Gallini (1984); Gallini and Winter (1985); Katz and Shapiro (1985a); Kamien and Tauman (1986); Gallini and Wright (1990); Anton and Yao (1994).[50]

Gallini (1984) emphasizes the strategic incentives for an incumbent to license its technology to an entrant to reduce the incentives for the entrant to do R&D;[51]. In a different vein, Gallini and Winter (1985), investigate incentives to license in a duopoly and its effect on R&D incentives. In their model firms always license but the availability of licensing can have differential effects on R&D effort depending on how competitive is the initial position of the two firms (measured in terms of the closeness of their production techniques). Licensing encourages R&D when firms are initially close but discourages it when they are asymmetric.[52] The authors also make the point that, where it is possible to keep information secret, patents may be seen as facilitating (rather than reducing) information flow since providing 'property-rights style' protection enables licensing. This argument is usually known under the title of the 'contract theory of patents' and has continued to receive attention in the literature.[53]

Meanwhile, Katz and Shapiro (1985b), investigate all of the main licensing questions using a three stage game where R&D is followed by licensing and then competition in the output market a la Cournot. In their model not all innovations are licensed, with low value innovations more likely to be licensed than high value ones. Regarding both R&D incentives and welfare the effect of licensing is ambiguous with a negative or positive impact possible depending on parameters.[54]

Kamien and Tauman (1986), look at the structure of licensing contracts where competition takes the Stackelberg form (with the innovator the leader). Non-drastic

[50]Closely related to the issue of licensing is the possibility of cooperation in R&D. In this literature spillovers play a prominent role (with cooperation being one means of internalising them). There is now a large literature, which we are not able cover in detail here – for examples see, Katz (1986); D'Aspremont and Jacquemin (1988); Katz and Ordover (1990); Suzumura (1992); Simpson and Vonortas (1994); Ziss (1994); Leahy and Neary (1997); Katsoulacos and Ulph (1998); Goyal and Moraga-Gonzalez (2001).

[51]A similar point is considered by Rockett (1990) (and following her (Eswaran, 1994)), who investigates selective licensing by incumbents as a strategic tool for 'selecting the competition' so as to prolong their dominance post patent expiry.

[52]This result has interesting analogies with the recent paper of Cabral and Polak (2007) who examine the relationship between dominance, imitation and innovation. They find that dominance is bad for R&D when intellectual property rights are weak but good when they are strong.

[53]For a recent example see Denicolo and Franzoni (2003). However, it should be noted that intellectual property rights are not essential to licensing knowledge even in the presence of the sorts of informational asymmetries emphasized by Arrow (1962) – see, for example, Anton and Yao (1994, 2002), who explore how an innovator might be able to extract rents under licensing even in the absence of intellectual property protection.

[54]The same authors produced a whole series of further papers on this topic, see for example Katz and Shapiro (1986) which examines the strategy of a research lab licensing to firms who are product market competitors, and Katz and Shapiro (1987) which examines the innovation effort in a duopoly when ex-post dissemination either via licensing or imitation is possible.

innovations are licensed to all competitors using a fixed fee (not a per-unit royalty) while drastic innovations are licensed to a single firm. Similar questions are addressed by Gallini and Wright (1990), who investigate the structure of licensing contracts (linear vs. non-linear, exclusive vs. non-exclusive) in the presence of asymmetric information and the possibility of imitation. They show that high and low value innovations will be licensed differently with low value innovations licensed exclusively for a fixed fee but high value innovations will usually use an output-based format (though possibly with a fixed fee).

2.6 Imitation[55]

The empirical literature on innovation and intellectual property, from an early stage, indicated that an intellectual property right, such as a patent, provided a very imperfect monopoly, with competing firms often able to 'imitate' a given innovation well before the formal expiry of the patent.[56] The same literature also tended to show that 'imitation' was a non-trivial exercise which even in the absence of a patent might require substantial time and effort.[57] This stood in contrast to much of the early theoretical literature, which as Levin (1986) emphasized, tended to assume that patents provided perfect excludability (and even in some cases perfect appropriability).

There were of course exceptions. Reinganum (1982),[58] incorporated the possibility of imitation (though in her model imitation simply yields a lower return to innovation – it is still costless and instantaneous). Horstmann, MacDonald, and Slivinski (1985), develop a model that allows imitation even where patents are present in an effort to explain why firms only patent a proportion of their innovations. Here a patent may signal to a competitor that opportunities are 'good' and hence encourage imitation (without the signal the competitor might simply exit the market leaving the innovator as the monopolist). On a different tack, Benoit (1985), has a duopoly model in which the innovation is not patentable and imitation by the non-innovator is possible. Here imitation is as costly as innovation but there is uncertainty about the value of an innovation which is only resolved once it is discovered. As a result, imitation may

[55]Closely related to the question of imitation is that of the diffusion of a given innovation. There is now a large literature on this topic which we will cover in detail in this review. As a starting point the reader is directed to Griliches (1957); Reinganum (1981); Jovanovic (1982); Jovanovic and MacDonald (1994); Gort and Klepper (1982); Klepper and Simons (2000).

[56]See for example Mansfield (1961); Taylor and Silberston (1973); Mansfield, Schwartz, and Wagner (1981); Mansfield (1985) and Levin et al. (1987).

[57]A point made particularly strongly by Dosi (1988).

[58]An even earlier example, that builds on the analysis of Scherer (1967) was Baldwin and Childs (1969). Another early work that included imitation to some extent was Futia (1980), who has an exogenous level of entry and imitation in his model of Schumpeterian competition.

drive down innovator rents: the innovator still loses on 'bad' innovations but now has its profits reduced on good ones; and, as a result, the level of innovation may be a non-monotonic function of innovation cost with a firm more willing to undertake higher cost innovations.[59]

Following this early work came the literature on patent breadth which we have already discussed above. Here, the relationship of patent scope had a direct impact on the development of imitative products. This was a substantial improvement in realism – there was now an explicit product space in which imitation did not have to be perfectly duplicative – but there was a tendency to still see patents as perfectly exclusionary within their scope and for imitation to be costless.[60] One paper that does allow for both costly imitation and product differentiation, though restricted by an assumption of exogenous participation (there are just two firms), is Harter (1994). Building on the model of R&D in a Hotelling model of product differentiation developed in Harter (1993), he allows patenting by the innovator and for imitation of the innovation (the effect of a patent here is two-fold: it makes imitation cheaper but the imitator must locate her product outside of the exclusion zone set by the patent).[61]

2.6.1 Endogenous Growth Style Models

A rather different strand of literature on the topic is that coming from quality-ladder style models of endogenous growth. These naturally tend to have a strong connection to the work on cumulative innovation already discussed.[62] Early work incorporating imitation in a dynamic general-equilibrium framework included that of Segerstrom (1991); Grossman and Helpman (1991b) and Helpman (1993).

Segerstrom (1991) and Grossman and Helpman (1991b) build similar models based on the framework developed in Grossman and Helpman (1991a) but allow firms to engage in costly imitation as well as innovation. Grossman and Helpman analyze

[59]Taken to an extreme, if imitation is sufficiently cheap and effective then firms will prefer to imitate rather than innovate and there will be a 'waiting-game' rather than a patent race – see Katz and Shapiro (1987); Dasgupta (1988) for early discussion and Choi (1998) who as part of a wider paper on patent-litigation, patent strength and imitation investigates waiting-game style behaviour in *imitation*.

[60]Such generalisations are never entirely accurate. Gallini (1992) has costly imitation though her model does not feature product differentiation.

[61]An example of an alternative approach where imitation costs are non-zero is that of Pepall and Richards (1994). Their model features quality choice by the innovator, uncertainty about demand, perfect but costly imitation and Stackelberg quantity competition in the final product market. They find that imitation may lead to welfare losses due to inefficiently low choice of product quality by the innovator.

[62]Indeed many of those models, particularly those with multiple stages, incorporate imitation. However they usually do so in a rather basic form – imitation is instantaneous outside of the scope of the patent and impossible within it.

a model with two regions: a 'North' and a 'South'. Innovation only takes place in the North and imitation only takes place in the South and in both cases follows a classic patent race form. Due to factor price differences if firms from both regions simultaneously have access to the same product quality the Southern firm produces (Bertrand competition with cost differences). By contrast, Segerstrom has a single region and a firm in a given industry may engage in both imitation and innovation that follow a patent-race format with imitation being cheaper than innovation, and firms with the same technology play an oligopoly game which allows for collusion (collusion does in fact occur in equilibrium yielding non-zero profits for firms even when there are multiple participants). As a result both imitation and innovation occur (though not at all stages) and imitation reduces incentives to innovate.[63]

Neither of these models explicitly considered the impact of intellectual property rights. This is something considered by Helpman (1993). However, the paper's focus is a rather 'macro' one, aimed at evaluating the different channels by which an increase in the strength of intellectual property rights impacts on welfare – whether via terms of trade, production composition, available products, intertemporal allocation of consumption, etc. As a result, the model of innovation and imitation is highly simplified.[64]

More sophisticated, recent, work that incorporates both competition and some measure of intellectual property protection is that of Aghion, Harris, Howitt, and Vickers (2001). Their model modifies the standard quality-ladder by turning each industry into a duopoly with differentiated products (so firms compete via price competition but not in a pure Bertrand form). Firms innovate to reduce costs and the laggard engages in imitation. Both processes take patent-race form but imitation occurs more quickly that innovation. Intellectual property rights are not explicitly present but there is general 'substitutability' parameter α which can be seen as proxying the absence of barriers (such as intellectual property) to direct entry into a rival firm's market.[65]

To conclude, all of these endogenous-growth style models offer a rich approach

[63]In a subsequent paper Segerstrom with co-author Davidson (Davidson and Segerstrom, 1998), investigate the impact of R&D subsidies on growth, in a similar endogenous-growth style model that again includes imitation as well as innovation.

[64]Helpman uses a North/South model similar to that of Krugman (1979) and imitation simply occurs at some exogenous rate determined by the intellectual property policy parameter. Neither innovation nor imitation require resources. This is obviously a substantial simplification (as Helpman acknowledges see fn. 5 p. 1250) but is in accord with the focus of the analysis.

[65]Aghion, along with co-authors, has done a substantial amount of subsequent work along similar lines. For example, Aghion, Bloom, Blundell, Griffith, and Howitt (2005) looks at how incorporating the level of product market competition into a Schumpeterian model can help explain the empirical finding of the 'inverted-U' shape relationship between innovation and competition observed empirically. However these papers tend to have a highly simplified model of imitation as their primary focus is elsewhere.

to considering imitation rather different from that found in a 'normal' IO literature. However, partially as a consequence of their complexity in other areas, they tend to be rather restrictive in two important ways. First, one cannot use them to explore inter-industry heterogeneity in innovation and imitation behaviour. Second, and most importantly, in contrast to the 'cumulative innovation' literature there is no modelling of micro modelling of the licensing process (an innovator or imitator never has to negotiate with existing producers).

2.6.2 Capital-Style Models of Free Replication of Knowledge

The effect of removing the assumption that imitation is instantaneous has been addressed, albeit using a rather more macroeconomic approach, with very interesting results in the recent work of Boldrin and Levine (2003, 2005) and Quah (2002) (hereafter BLQ). In these models 'ideas' are treated like capital in a standard macroeconomic general equilibrium model, and, once created, have a standard neoclassical production function determining the rate at which new copies can be made. The main difference between 'ideas' and capital is that there is a one-off charge to create the first 'copy' of an 'idea' (the fixed cost of the innovator). In equilibrium, if the 'idea' is to be instantiated, this fixed cost must be less than the first period price (the income received by the innovator). It is shown (the most thorough treatment is by Quah) that, in the absence of intellectual property rights (i.e. under conditions of free competition): a) initial prices are bounded away from zero and thus the level of innovation is non-zero b) (Quah Thm 4.9) that there exists a non-trivial competitive innovation equilibrium c) (Quah 4.10 and Fig. 1) this equilibrium will (probably) not be socially efficient (i.e. there are conditions under which it will be efficient but these conditions are rather restrictive) d) changing the rate of reproducibility, that is the rate at which one can copy, may increase the first period price and therefore the revenue to the creator of the first copy.

BLQ are making an important point in highlighting the restrictive nature of a pure nonrivalry assumption. However, there are, in turn, several problems with their alternative. Most fundamentally, while it is undoubtedly true that new 'ideas' must be embodied, be it in goods, services or human capital in order to be useful this does not necessarily make the underlying 'ideas' nonrival. Suppose, for example (following Romer (1990)), that we have a new design for a hard disk drive which halves the per unit storage cost. Now, while it is clear that only the disk drives themselves have value to end consumers, nevertheless if the design can be copied at less than the cost of its original development we still have all the traditional problems: competition will drive price to marginal cost of production plus the cost of copying the new design and,

assuming the cost of copying is less than the cost of the original development, the creator of the original design will make a loss.

BLQ's model avoids this outcome by equating idea production with capital production in standard neoclassical macroeconomic models. Just as new capital is produced from old in those models so new copies of an idea are made from old. But this analogy is misleading, since it papers over the fundamental distinction between capital in a neoclassical growth model and ideas in an innovation model: while reproduction of capital can be viewed as a homogeneous process (though even this might be dubious) reproduction of ideas is not. Once an imitator has made the initial copy of an idea, 'normal' production, using capital and labour, kicks in and there is no constant returns to scale in the idea. But if that is so, other than the delay (which *is* important and is the major insight of these models), we are back to our original situation where the original innovator will be out of pocket.

In explicit production function terms: if any copy can be used as a basis for reproduction – as in BLQ, but that, unlike BLQ, once one copy of an 'idea' is made you can make additional ones using capital following a CRS production function $f(n, k)$ where n is the number of ideas (think of reproducing CDs be it as stamped plastic in a factory or as bits on a computer) then: $f(0, k) = 0, f(n, k) = f(1, k)$ for all $n \neq 0$ and $f(1, k) = \alpha k$. Thus, there is nonconvexity with respect to ideas. Under competition this implies that any second period price must be α and profits are zero. But then no-one would be willing to pay more than 0 for a copy of the idea and the originator cannot cover development costs.

Nonetheless BLQ do perform a valuable service in focusing attention on the fact that reproduction is not instantaneous. This ties in closely with the empirical fact of lead-time advantages. However to understand this fully we must introduce a clear distinction between imitation and reproduction. Imitation is the making of a first copy – a template – by a new producer who is not the originator. Once a producer has this first copy it may engage in reproduction: the making copies of its own copy in a standard manner.

Armed with this definition traditional nonrivalry can now be interpreted as the assumption that imitation is the same as reproduction. Conversely, with this definition, it is easy to see the similarities of imitation to original innovation:

1. A fixed cost of creating a first 'copy': imitators have 'development' costs just like innovators.

2. Producing a 'copy' takes time: imitation just like innovation is not instantaneous.

2.7 Open[66] Approaches to Knowledge Production[67]

Recent years have seen a variety of areas in which open approaches to knowledge production feature prominently. For example, in the software industry we have the phenomenon of open source software[68] while in the area of online content we have sites such as Wikipedia.[69] Such developments stimulate one to ask: how well can an open approach to knowledge production do? Are there models in which an open approach to knowledge production would be optimal. In particular, how (and why) could an open approach to the production of knowledge goods be superior, in terms of innovative output, to one based on exclusive rights? It is important to note here that we are focused on the rate of innovation and not the level of welfare. After all it is well-accepted that being more 'open' (having weaker intellectual property rights) can improve welfare by improving access.[70] But this is certainly not the case in relation to innovation. In fact most of the literature, implicitly or explicitly, would support the following propositions:

Proposition 2.7.1. *The level of R&D (and hence the rate of innovation) is increasing in the payoff from successful R&D (e.g. the level of reward from winning a patent race).*

Proposition 2.7.2. *Strengthening intellectual property rights such as patents increase the payoff to successful R&D (e.g. a patent is more valuable if it covers more or lasts for longer).*

Corollary 2.7.3. *The rate of innovation is a monotonically increasing in the level of*

[66] An 'open' approach to knowledge production is one where the resulting knowledge is 'open', that is, it can be freely used, redistributed and reused. The word 'freely' must be loosely interpreted – for example the requirement of attribution or even that derivative works be re-shared, does not render a work unfree. However it does exclude the requirement of payment, or the imposition or restrictions on the type of use (such as limiting the use to research or non-commercial activities). Furthermore, since, without access, a piece of knowledge cannot be used it also excludes the use of secrecy – 'open' knowledge must be publicly available.

[67] The discussion in this section can usefully be supplemented by the more extensive survey in Pollock (2006a).

[68] The literature on open-source is growing rapidly. For an introduction and overview see Lerner and Tirole (2002, 2005); Maurer and Scotchmer (2006). Examples of early work include Benkler (2002); Von Hippel (2002); Casadesus-Masanell and Ghemawat (2003); Lakhani and von Hippel (2003); Gaudeul (2004); Bonaccorsi and Rossi (2004); Bessen (2006).

[69] Of course, open approaches are by no means new: consider the two century old example of John Rennie, one of the most famous engineers of the industrial revolution. In 1789 he worked on the Albion Mills for Watt and Boulton. To Watt's horror, upon completion, Rennie, rather than patenting his new design, was eager to demonstrate it to others. "[F]ar from ruining him [Rennie] as Watt predicted, [this] established his reputation and led to a flood of commissions"(Macleod, 1988, p. 104). Nevertheless the increasing prominence of 'knowledge' in the economy has brought these questions a new prominence and significance.

[70] This assumption is implicit in the literature on the subject of optimal patent design for, without it, in most of those models optimal patents would be infinitely long and broad.

intellectual property rights, that is strengthening the degree of protection (and therefore increasing the reward for a winning firm) always increases the rate of innovation.

Thus, in order for an open approach to be a better *production* model requires us to identify where one or other of the above propositions is in error.

2.7.1 Innovating Theory

Given the innate plausability of the first of the two propositions our focus must be on the second. In particular one can consider the two ends of the production equation: one can investigate (a) whether intellectual property imposes costs that openness does not and/or (b) whether the discrepancy in incentives (monetary or otherwise) between an open regime and an intellectual property regime is less substantial than initially imagined (in the crudest models, where pure nonrivalry is assumed, income for innovators is zero in the absence of intellectual property rights).[71]

On the cost side there are various points to be made. With cumulative innovation the rights of new innovators may overlap with those of old. Combined with obstacles to perfectly efficient bargaining (such as imperfect information) exclusive rights may result in hold-up. This approach appears in both Bessen and Maskin (2006) (discussed above in Section 2.4) and Pollock (2006b). Both papers find that, in certain circumstances, it will be preferable to have an open, rather than an intellectual property rights, regime.[72]

One would also expect the level of componentization to play a role (for example, one would expect the degree of hold-up to increase with the level of 'componentization'. Componentization is used as a generic term here to denote the situation where a given product or idea combines or depends upon many previous ones (rather than a single one). As yet, there are very few papers that address the question of innovation, and innovation policy, in the area of componentized goods (and none that the author knows of which address componentized *and* cumulative innovation). Shapiro (2001) considers cross-licensing and patents pools and makes the general point that pools improve welfare when the patents are complements but harm welfare when the patents are substitutes. Lerner and Tirole (2004) develop a more complex model of patent pool arrangements seeking to provide some general guidelines as to when such pools are welfare improving. Meanwhile, Gilbert and Katz (2007), develop a patent

[71]Such an assumption is equivalent to assuming instantaneous and costless imitation. Such an assumption, which is a natural one to make when focusing on other issues, is pervasive in the literature – appearing explicitly for example in Klemperer (1990); Hopenhayn and Mitchell (2001); Menell and Scotchmer (2005).

[72]See also the model of Hunt (2006) who develops a simple model in which patents may reduce R&D.

race model for 'complex' technologies (those with many components) and investigate what the optimal division of profit should be in order to induce efficient R&D effort.[73]

A second option, also related to the cumulativeness of innovation, is that participation (production) at different innovation stages are linked – for example, one could have that participation at any given stage in the innovation 'ladder' is dependent on participation at the previous stage. In this case, intellectual property rights, by excluding innovators from participation at stage N, reduce those who can participate at future stages. In such a situation it is possible that all innovators lose out in the long run – even those who, by successfully obtaining intellectual property rights, gain in the short term. As a result the level of innovation will be reduced compared to the situation without intellectual property rights.

Turning to income side of the equation the first point to consider is the possibility of up-front funding. With up-front funding either by rewards or by direct subsidy (research in universities for example) it is possible for work to be open ab initio *and*, at the same time, for their 'creators' to be guaranteed remuneration.[74]

Even without up-front funding it is often possible for creators to derive a substantial income by means other than by the use of exclusive rights. Of course, one must be careful here since the primary alternative to the use of intellectual property is not openness but secrecy. Thus, in considering the various methods by which remuneration can be obtained, we should confine ourselves to those mechanisms that are compatible with open production (that is those which ensure the knowledge produced is 'open').

The most prominent examples of such mechanisms arise where there exists a rival good which is complementary to the underlying knowledge.[75] Examples of such a complementary rival good include support services in relation to open source software, live performances in relation to 'open' music,[76] and access to attention in the case of advertising supported information provision.

We should also add a qualification to the implicit assumption of opposition between

[73]The componentization of production in an industry combined with the presence of intellectual property rights can lead to patent 'thickets' which obstruct innovation. Hall and Ziedonis (2001) provide evidence for this effect in the semiconductor industry while Bessen and Hunt (2007) do so for the software industry – in this area there is also the recent work of Noel and Schankerman (2006) which looks at the overall effect of patents in the software industry (focusing on large firms only) and while finding some negative impact of 'thicketiness' find an overall positive impact of patents on R&D.

[74]OECD (2005) figures indicate that in 2004 private firms accounted for approximately 53% of total expenditure R&D with the remainder coming from public sources. In the USA and Japan the private share is higher at 63% and 74% respectively. In Latin American by contrast the public share is the majority (NSF 2000). For work on alternative compensation systems and 'prize design' see Wright (1983); De Laat (1996); Kremer (1998); Shavell and van Ypersele (2001); Fisher (2004).

[75]The potential use of complementary goods as an alternative method of appropriation when intellectual property is weak found particular emphasis in the seminal article of Teece (1986) – revisited in a recent special issue of Research Policy(volume 35, number 8, October 2006).

[76]See, for example, Connolly and Krueger (2005); Mortimer and Sorenson (2005).

openness and intellectual property. It should be remembered that the relationship between the open 'commons' and the enclosed realm of intellectual property rights is not a purely antagonistic one. As intellectual property rights expire, the knowledge they cover flows into the public domain, increasing and enriching it. Conversely, it is a fact universally acknowledged that all creators must be in want of a rich and vibrant public domain on which to build and from which to derive new ideas. Of course, the history of intellectual property, or at least copyright, can provide many instances where this flow has been dammed or even reversed by sudden expansions in the scope or duration of rights (or even where such changes are usually applied equally to existing and prospective work thereby removing work from the public domain). Nevertheless the fact remains that, at least when not abused, the relationship can be a symbiotic one rather than one of rivalry and opposition.

2.7.2 Conclusion

From the above summary it should be clear that there are indeed reasons why the propositions, and their associated corollary, might fail, and for 'openness' to be good for innovation. That said, whilst progress is being made, there is, as yet, no fully articulated and intellectually coherent theory, or empirics, of open knowledge production that can convincingly demonstrate its advantages when compared to other approaches, such as those based on exclusive rights (intellectual property).

Furthermore, it is necessary to go beyond simple explanation, to examine in detail both (a) the various factors at work that influence the attractiveness (or not) of an open approach and (b) how these factors relate to the different types of subject matter. Is it, for example, the feasability of up-front funding, the presence of strong first-mover advantages, the level of transaction costs or the degree of componentization – among many other factors – that determine the advantages (and disadvantages) of an open approach vis-a-vis intellectual property? And are these factors constant or do they vary across disciplines? Are the same factors equally important in the production of pharmaceuticals and the development of operating systems – or, for that matter, online encyclopaedias? If not, as seems likely, then any general theory will need careful calibration to the specifics of the case at hand.

Bibliography

Philippe Aghion and Peter Howitt. A Model of Growth Through Creative Destruction. *Econometrica*, 60(2):323–351, March 1992. ISSN 00129682.

Philippe Aghion, Christopher Harris, Peter Howitt, and John Vickers. Competition, Imitation and Growth with Step-by-Step Innovation. *The Review of Economic Studies*, 68(3):467–492, July 2001. ISSN 00346527.

Philippe Aghion, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. Competition and Innovation: An Inverted-U Relationship. *The Quarterly Journal of Economics*, 120(2):701–728, May 2005.

Bharat N Anand and Tarun Khanna. The Structure of Licensing Contracts. *Journal of Industrial Economics*, 48(1):103–35, March 2000.

James J Anton and Dennis A Yao. Expropriation and Inventions: Appropriable Rents in the Absence of Property Rights. *The American Economic Review*, 84(1):190–209, March 1994. ISSN 00028282.

James J Anton and Dennis A Yao. The Sale of Ideas: Strategic Disclosure, Property Rights, and Contracting. *The Review of Economic Studies*, 69(3):513–531, July 2002. ISSN 00346527.

Kenneth Arrow. Economic Welfare and the Allocation of Resources for Invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pages 619–625. Princeton University Press, 1962.

Kenneth Arrow. The Interaction of Corporate Market Allocation Processes and Entrepreneurial Activity. In R. H. Day, G. Eliasson, and C. Wihlborg, editors, *The Markets for Innovation, Ownership and Control*. Amsterdam: North-Holland, 1993.

William L Baldwin and Gerald L Childs. The Fast Second and Rivalry in Research and Development. *Southern Economic Journal*, 36(1):18–24, July 1969. ISSN 00384038.

Yoram Barzel. Optimal Timing of Innovations. *The Review of Economics and Statistics*, 50(3):348–355, August 1968. ISSN 00346535.

John Beath, Yannis Katsoulacos, and David Ulph. Sequential Product Innovation and Industry Revolution. *The Economic Journal*, 97(Supplement: Conference Papers): 32–43, 1987. ISSN 00130133.

Yochai Benkler. Coase's Penguin, or Linux and The Nature of the Firm. *Yale Law Journal*, 112(3):369–446, 2002.

Jean-Pierre Benoit. Innovation and Imitation in a Duopoly. *The Review of Economic Studies*, 52(1):99–106, 1985. ISSN 00346527.

James Bessen. Hold-up and Patent Licensing of Cumulative Innovations with Private Information. *Economics Letters*, 82(3):321–326, 2004.

James Bessen. Open Source Software: Free Provision of Complex Public Goods. In Jargen Bitzer and Philipp J. H. Schraeder, editors, *The Economics of Open Source Software Development*. Elsevier B. V., 2006.

James Bessen and Robert M. Hunt. An Empirical Look at Software Patents. *Journal of Economics & Management Strategy*, 16(1):157–189, 03 2007.

James Bessen and Eric Maskin. Sequential Innovation, Patents, and Innovation. NajEcon Working Paper Reviews 321307000000000021, www.najecon.org, May 2006.

Sudipto Bhattacharya and Dilip Mookherjee. Portfolio Choice in Research and Development. *The RAND Journal of Economics*, 17(4):594–605, 1986. ISSN 07416261.

Michele Boldrin and David Levine. Perfectly Competitive Innovation, 1 2003. Unpublished working paper. First version 1997-10-03.

Michele Boldrin and David Levine. IP and market size. Levine's Working Paper Archive 618897000000000836, UCLA Department of Economics, July 2005.

A. Bonaccorsi and C. Rossi. Altruistic individuals, selfish firms? The structure of motivation in Open Source software. *First Monday*, 9(1), 2004.

Christopher Budd, Christopher Harris, and John Vickers. A Model of the Evolution of Duopoly: Does the Asymmetry between Firms Tend to Increase or Decrease? *The Review of Economic Studies*, 60(3):543–573, July 1993. ISSN 00346527.

Harun Bulut and GianCarlo Moschini. Patents, trade secrets and the correlation among R&D projects. *Economics Letters*, 91(1):131–137, April 2006.

Luis Cabral. Bias in market R&D portfolios. *International Journal of Industrial Organization*, 12(4):533–547, December 1994.

Luis Cabral and Ben Polak. Dominant Firms, Imitation, and Incentives to Innovate. Working Papers 07-6, New York University, Leonard N. Stern School of Business, Department of Economics, 2007.

Ramon Casadesus-Masanell and Pankaj Ghemawat. Dynamic mixed duopoly: A model motivated by Linux vs. Windows. IESE Research Papers D/519, IESE Business School, September 2003. Forthcoming in Management Science.

Howard F Chang. Patent Scope, Antitrust Policy, and Cumulative Innovation. *The RAND Journal of Economics*, 26(1):34–57, 1995. ISSN 07416261.

Shubham Chaudhuri, Pinelopi K. Goldberg, and Panle Jia. Estimating the Effects of Global Patent Protection in Pharmaceuticals: A Case Study of Quinolones in India. *The American Economic Review*, 96(5):1477–1514, December 2006. ISSN 0002-8282. doi: 10.1257/000282806779396111.

Jay Pil Choi. Patent Litigation as an Information-Transmission Mechanism. *The American Economic Review*, 88(5):1249–1263, December 1998. ISSN 00028282.

Marie Connolly and Alan Krueger. Rockonomics: The Economics of Popular Music. Working Papers 499, Princeton University, Department of Economics, Industrial Relations Section., April 2005.

Francesca Cornelli and Mark Schankerman. Patent Renewals and R&D Incentives. *The RAND Journal of Economics*, 30(2):197–213, 1999. ISSN 07416261.

Partha Dasgupta. Patents, Priority and Imitation or, the Economics of Races and Waiting Games. *The Economic Journal*, 98(389):66–80, March 1988. ISSN 00130133.

Partha Dasgupta. The Economics of Parallel Research. In Frank Hahn, editor, *The Economic Theory of Information, Games, and Missing Markets*. Oxford University Press, 1989.

Partha Dasgupta and Eric Maskin. The Simple Economics of Research Portfolios. *The Economic Journal*, 97(387):581–595, 1987.

Partha Dasgupta and Joseph Stiglitz. Uncertainty, Industrial Structure, and the Speed of R&D. *The Bell Journal of Economics*, 11(1):1–28, 1980a.

Partha Dasgupta and Joseph Stiglitz. Industrial Structure and the Nature of Innovative Activity. *The Economic Journal*, 90(358):266–293, June 1980b. ISSN 00130133.

Claude D'Aspremont and Alexis Jacquemin. Cooperative and Noncooperative R & D in Duopoly with Spillovers. *The American Economic Review*, 78(5):1133–1137, December 1988. ISSN 00028282.

Carl Davidson and Paul Segerstrom. R&D Subsidies and Economic Growth. *The RAND Journal of Economics*, 29(3):548–577, 1998. ISSN 07416261.

E. A. De Laat. Patents or Prizes: Monopolistic R&D and Asymmetric Information. *International Journal of Industrial Organization*, 15:369–390, 1996.

Flavio Delbono. Market Leadership With a Sequence of History Dependent Patent Races. *The Journal of Industrial Economics*, 38(1):95–101, September 1989. ISSN 00221821.

Flavio Delbono and Vincenzo Denicolo. Incentives to Innovate in a Cournot Oligopoly. *The Quarterly Journal of Economics*, 106(3):951–961, August 1991. ISSN 00335533.

Victor Denicolo. Two-Stage Patent Races and Patent Policy. *Rand Journal of Economics*, 31:488–501, 2000.

Vincenzo Denicolo. Patent Races and Optimal Patent Breadth and Length. *The Journal of Industrial Economics*, 44(3):249–265, September 1996. ISSN 00221821.

Vincenzo Denicolo and Luigi Alberto Franzoni. The Contract Theory of Patents. *International Review of Law and Economics*, 23(4):365–380, December 2003.

Avinish Dixit. A General Model of R&D Competition and Policy. *The RAND Journal of Economics*, 19(3):317–326, 1988.

Giovanni Dosi. Sources, Procedures, and Microeconomic Effects of Innovation. *Journal of Economic Literature*, pages 1120–1171, 1988.

Mukesh Eswaran. Licensees as Entry Barriers. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 27(3):673–688, August 1994. ISSN 00084085.

Joseph Farrell, Richard Gilbert, and Michael Katz. Market Structure, Organizational Structure, and R&D Diversity. Department of Economics, Working Paper Series 1049, Department of Economics, Institute for Business and Economic Research, UC Berkeley, October 2002.

William W. Fisher. *Promises to Keep: Technology, Law, and the Future of Entertainment*. Stanford University Press, August 2004. ISBN 0804750130.

Dominique Foray. *The Economics of Knowledge*. MIT, 2004.

Drew Fudenberg, Richard Gilbert, Joseph Stiglitz, and Jean Tirole. Preemption, leapfrogging and competition in patent races. *European Economic Review*, 22(1): 3–31, June 1983.

Carl A Futia. Schumpeterian Competition. *The Quarterly Journal of Economics*, 94 (4):675–695, June 1980. ISSN 00335533.

J. J. Gabszewicz and J.-F. Thisse. Entry (and Exit) in a Differentiated Industry. *Journal of Economic Theory*, 22:327–338, 1980.

Nancy Gallini. Patent Policy and Costly Imitation. *Rand Journal of Economics*, 23 (1):52–63, 1992.

Nancy T Gallini. Deterrence by Market Sharing: A Strategic Incentive for Licensing. *The American Economic Review*, 74(5):931–941, December 1984. ISSN 00028282.

Nancy T Gallini and Ralph A Winter. Licensing in the Theory of Innovation. *The RAND Journal of Economics*, 16(2):237–252, 1985. ISSN 07416261.

Nancy T Gallini and Brian D Wright. Technology Transfer under Asymmetric Information. *The RAND Journal of Economics*, 21(1):147–160, 1990. ISSN 07416261.

Alexandre Gaudeul. Open Source Software Development Patterns and License Terms. Technical report, September 2004.

Richard Gilbert and Michael Katz. Efficient Division of Profit for Complex Technologies, 2007. Unpublished working paper.

Richard Gilbert and Carl Shapiro. Optimal Patent Length and Breadth. *The RAND Journal of Economics*, 21(1):106–112, 1990. ISSN 07416261.

Richard J Gilbert and David M. G Newbery. Preemptive Patenting and the Persistence of Monopoly. *The American Economic Review*, 72(3):514–526, June 1982. ISSN 00028282.

Michael Gort and Steven Klepper. Time Paths in the Diffusion of Product Innovations. *Economic Journal*, 92(367):630–653, 1982.

Sanjeev Goyal and Jose Luis Moraga-Gonzalez. R&D Networks. *The RAND Journal of Economics*, 32(4):686–707, 2001. ISSN 07416261.

Jerry Green and Suzanne Scotchmer. On the Division of Profit between Sequential Innovators. *Rand Journal of Economics*, 26(1):20–33, 1995.

Zvi Griliches. Hybrid Corn: An Exploration in the Economics of Technological Change. *Econometrica*, 25(4):501–522, October 1957. ISSN 00129682.

Gene M Grossman and Elhanan Helpman. Quality Ladders in the Theory of Growth. *The Review of Economic Studies*, 58(1):43–61, 1991a. ISSN 00346527.

Gene M Grossman and Elhanan Helpman. Quality Ladders and Product Cycles. *The Quarterly Journal of Economics*, 106(2):557–586, May 1991b. ISSN 00335533.

Gene M Grossman and Carl Shapiro. Optimal Dynamic R&D Programs. *The RAND Journal of Economics*, 17(4):581–593, 1986. ISSN 07416261.

Gillian K. Hadfield. The Economics of Copyright: An Historical Perspective. *Copyright Law Symposium*, 38:1–46, 1992.

Bronwyn Hall and Rosemarie Ziedonis. The patent paradox revisited: an empirical study of patenting in the U.S. semiconductor industry, 1979-1995. *Rand Journal of Economics*, 32(1):101–128, 2001.

Arnold C Harberger. Monopoly and Resource Allocation. *The American Economic Review*, 44(2, Papers and Proceedings of the Sixty-sixth Annual Meeting of the American Economic Association):77–87, May 1954. ISSN 00028282.

Christopher Harris and John Vickers. Perfect Equilibrium in a Model of a Race. *Review of Economic Studies*, 52(2):193–209, April 1985a.

Christopher Harris and John Vickers. Patent Races and the Persistence of Monopoly. *The Journal of Industrial Economics*, 33(4, A Symposium on Oligopoly, Competition and Welfare):461–481, June 1985b. ISSN 00221821.

Christopher Harris and John Vickers. Racing with Uncertainty. *The Review of Economic Studies*, 54(1):1–21, 1987. ISSN 00346527.

John F. R Harter. Differentiated Products with R&D. *The Journal of Industrial Economics*, 41(1):19–28, March 1993. ISSN 00221821.

John F. R Harter. The Propensity to Patent with Differentiated Products. *Southern Economic Journal*, 61(1):195–201, July 1994. ISSN 00384038.

Elhanan Helpman. Innovation, Imitation, and Intellectual Property Rights. *Econometrica*, 61(6):1247–1280, November 1993. ISSN 00129682.

Hugo A Hopenhayn and Matthew F Mitchell. Innovation Variety and Patent Breadth. *The RAND Journal of Economics*, 32(1):152–166, 2001. ISSN 07416261.

Andrew W Horowitz and Edwin L. C Lai. Patent Length and the Rate of Innovation. *International Economic Review*, 37(4):785–801, November 1996. ISSN 00206598.

Ira Horowitz. Research Inclinations of a Cournot Oligopolist. *The Review of Economic Studies*, 30(2):128–130, June 1963. ISSN 00346527.

Ignatius Horstmann, Glenn M MacDonald, and Alan Slivinski. Patents as Information Transfer Mechanisms: To Patent or (Maybe) Not to Patent. *The Journal of Political Economy*, 93(5):837–858, October 1985. ISSN 00223808.

Robert M. Hunt. Patentability, Industry Structure, and Innovation. *Journal of Industrial Economics*, 52(3):401–425, 09 2004.

Robert M. Hunt. When Do More Patents Reduce R&D? *American Economic Review*, 96(2):87–91, May 2006.

R. Hurt and R. Schuchman. The Economic Rationale for Copyright. *American Economic Review*, 56(1):421–432, 1966.

Thomas Jefferson. *The Writings of Thomas Jefferson*. Washington: Thomas Jefferson Memorial Association, 1905.

Boyan Jovanovic. Selection and the Evolution of Industry. *Econometrica*, 50(3):649–670, May 1982. ISSN 00129682.

Boyan Jovanovic and Glenn M MacDonald. Competitive Diffusion. *The Journal of Political Economy*, 102(1):24–52, February 1994. ISSN 00223808.

Kenneth L Judd. On the Performance of Patents. *Econometrica*, 53(3):567–586, May 1985. ISSN 00129682.

M. Kamien and N. Schwartz. *Market Structure and Innovation*. CUP, 1982.

Morton I Kamien and Nancy L Schwartz. Timing of Innovations Under Rivalry. *Econometrica*, 40(1):43–60, 1972a. ISSN 00129682.

Morton I Kamien and Nancy L Schwartz. Market Structure, Rivals' Response, and the Firm's Rate of Product Improvement. *The Journal of Industrial Economics*, 20(2): 159–172, April 1972b. ISSN 00221821.

Morton I Kamien and Nancy L Schwartz. Patent Life and R&D Rivalry. *The American Economic Review*, 64(1):183–187, March 1974. ISSN 00028282.

Morton I Kamien and Nancy L Schwartz. Potential Rivalry, Monopoly Profits and the Pace of Inventive Activity. *The Review of Economic Studies*, 45(3):547–557, October 1978. ISSN 00346527.

Morton I Kamien and Yair Tauman. Fees Versus Royalties and the Private Value of a Patent. *The Quarterly Journal of Economics*, 101(3):471–492, August 1986. ISSN 00335533.

Yannis Katsoulacos and David Ulph. Endogenous Spillovers and the Performance of Research Joint Ventures. *The Journal of Industrial Economics*, 46(3):333–357, September 1998. ISSN 00221821.

Michael Katz and Carl Shapiro. Network Externalities, Competition, and Compatibility. *American Economic Review*, 75:424–440, 1985a.

Michael L Katz. An Analysis of Cooperative Research and Development. *The RAND Journal of Economics*, 17(4), 1986. ISSN 07416261.

Michael L Katz and Janusz A Ordover. R&D Cooperation and Competition. *Brookings Papers on Economic Activity. Microeconomics*, 1990:137–203, 1990. ISSN 10578641.

Michael L Katz and Carl Shapiro. On the Licensing of Innovations. *The RAND Journal of Economics*, 16(4):504–520, 1985b. ISSN 07416261.

Michael L Katz and Carl Shapiro. How to License Intangible Property. *The Quarterly Journal of Economics*, 101(3):567–590, August 1986. ISSN 00335533.

Michael L Katz and Carl Shapiro. R&D Rivalry with Licensing or Imitation. *The American Economic Review*, 77(3):402–420, June 1987. ISSN 00028282.

Edmund W Kitch. The Nature and Function of the Patent System. *Journal of Law and Economics*, 20(2):265–290, October 1977. ISSN 00222186.

Paul Klemperer. How Broad Should the Scope of Patent Protection Be? *RAND Journal of Economics*, 21(1):113–130, 1990.

Steven Klepper and K. Simons. The Making of an Oligopoly: Firm Survival and Technological Change in the Evolution of the U.S. Tire Industry. *Journal of Political Economy*, 108(4):728–760, 2000.

Michael Kremer. Patent Buyouts: A Mechanism for Encouraging Innovation. *The Quarterly Journal of Economics*, 113(4):1137–1167, November 1998. ISSN 00335533.

Paul Krugman. A Model of Innovation, Technology Transfer, and the World Distribution of Income. *The Journal of Political Economy*, 87(2):253–266, April 1979. ISSN 00223808.

Karim R. Lakhani and Eric von Hippel. How open source software works: "free" user-to-user assistance. *Research Policy*, 32(6):923–943, 2003.

William Landes and Richard Posner. An Economic Analysis of Copyright Law. *Journal of Legal Studies*, 18(2):325–363, 1989.

Jean Olson Lanjouw. Patent Protection in the Shadow of Infringement: Simulation Estimations of Patent Value. *The Review of Economic Studies*, 65(4):671–710, October 1998. ISSN 00346527.

Dermot Leahy and J. Peter Neary. Public Policy Towards R&D in Oligopolistic Industries. *The American Economic Review*, 87(4):642–662, September 1997. ISSN 00028282.

Tom Lee and Louis L Wilde. Market Structure and Innovation: A Reformulation. *The Quarterly Journal of Economics*, 94(2):429–36, March 1980.

Josh Lerner. An Empirical Exploration of a Technology Race. *The RAND Journal of Economics*, 28(2):228–247, 1997. ISSN 07416261.

Josh Lerner and Jean Tirole. Some Simple Economics of Open. *Journal of Industrial Economics*, 50(2):197–234, June 2002.

Josh Lerner and Jean Tirole. Efficient Patent Pools. *The American Economic Review*, 94(3):691–711, June 2004. ISSN 00028282.

Josh Lerner and Jean Tirole. The Economics of Technology Sharing: Open Source and Beyond. *Journal of Economic Perspectives*, 19(2):99–120, Spring 2005.

Richard Levin, A. Klevorick, R. Nelson, S. Winter, R. Gilbert, and Z. Griliches. Appropriating the Returns from Industrial Research and Development. *Brookings Papers on Economic Activity*, 3:783–831, 1987.

Richard C Levin. A New Look at the Patent System. *The American Economic Review*, 76(2, Papers and Proceedings of the Ninety-Eighth Annual Meeting of the American Economic Association):199–202, May 1986. ISSN 00028282.

Glenn C Loury. Market Structure and Innovation. *The Quarterly Journal of Economics*, 93(3):395–410, August 1979.

C. Macleod. *Inventing the Industrial Revolution: The English Patent System, 1660-1800*. Cambridge, 1988.

Edwin Mansfield. Technical Change and the Rate of Imitation. *Econometrica*, 29(4): 741–766, October 1961. ISSN 00129682.

Edwin Mansfield. How Rapidly Does New Industrial Technology Leak Out? *Journal of Industrial Economics*, 34(2):217–223, 1985.

Edwin Mansfield, Mark Schwartz, and Samuel Wagner. Imitation Costs and Patents: An Empirical Study. *Economic Journal*, 91(364):907–918, 1981.

Carmen Matutes, Pierre Regibeau, and Katharine Rockett. Optimal Patent Design and the Diffusion of Innovations. *The RAND Journal of Economics*, 27(1):60–83, 1996. ISSN 07416261.

Stephen M. Maurer and Suzanne Scotchmer. Open Source Software: The New Intellectual Property Paradigm. NBER Working Papers 12148, National Bureau of Economic Research, Inc, April 2006.

Peter Menell and Suzanne Scotchmer. Intellectual Property, 6 2005. forthcoming, Handbook of Law and Economics. Mitch Polinsky and Steven Shavell, eds. Amsterdam: Elsevier.

Robert Merges and Richard Nelson. On the Complex Economics of Patent Scope. *Columbia Law Review*, 90:839–916, 5 1990.

Julie Mortimer and Alan Sorenson. Supply Responses to Digital Distribution: Recorded Music and Live Performances. Technical report, 2005.

Richard Nelson and Sidney Winter. *An Evolutionary Theory of Economic Change*. Harvard University Press, 1982.

Michael D. Noel and Mark Schankerman. Strategic Patenting and Software Innovation. CEPR Discussion Papers 5701, C.E.P.R. Discussion Papers, May 2006.

William Nordhaus. *Invention, Growth and Welfare: A Theoretical Treatment of Technological Change*. M.I.T. Press, 1969.

Ted O'Donoghue. A Patentability Requirement for Sequential Innovation. *The RAND Journal of Economics*, 29(4):654–679, 1998. ISSN 07416261.

Ted O'Donoghue, Suzanne Scotchmer, and Jacques Thisse. Patent Breadth, Patent Life, and the Pace of Technological Improvement. *Journal of Economics and Management Strategy*, 7:1–32, 1998.

Ariel Pakes. Patents as Options: Some Estimates of the Value of Holding European Patent Stocks. *Econometrica*, 54(4):755–784, July 1986. ISSN 00129682.

Lynne M Pepall and Daniel J Richards. Innovation, Imitation, and Social Welfare. *Southern Economic Journal*, 60(3):673–684, 1994. ISSN 00384038.

Arnold Plant. The Economic Theory Concerning Patents for Inventions. *Economica*, 1(1):30–51, 2 1934a.

Arnold Plant. The Economic Aspects of Copyright in Books. *Economica*, 1(2):167–195, 5 1934b.

Rufus Pollock. The Value of the Public Domain. Technical report, Institute for Public Policy Research, July 2006a.

Rufus Pollock. Cumulative Innovation, Sampling and the Hold-up Problem. DRUID Working Papers 06-29, DRUID, Copenhagen Business School, Department of Industrial Economics and Strategy/Aalborg University, Department of Business Studies, 2006b.

Daniel Quah. Almost Efficient Innovation By Pricing Ideas, June 2002. Unpublished LSE working paper.

Jennifer F Reinganum. On the Diffusion of New Technology: A Game Theoretic Approach. *The Review of Economic Studies*, 48(3):395–405, July 1981. ISSN 00346527.

Jennifer F Reinganum. A Dynamic Game of R&D: Patent Protection and Competitive Behavior. *Econometrica*, 50(3):671–88, May 1982.

Jennifer F Reinganum. Uncertain Innovation and the Persistence of Monopoly. *The American Economic Review*, 73(4):741–748, September 1983. ISSN 00028282.

Jennifer F Reinganum. Innovation and Industry Evolution. *The Quarterly Journal of Economics*, 100(1):81–99, February 1985. ISSN 00335533.

Katharine E Rockett. Choosing the Competition and Patent Licensing. *The RAND Journal of Economics*, 21(1):161–171, 1990. ISSN 07416261.

Paul Romer. Endogenous Technological Change. *Journal of Political Economy*, 98 (5(2)):S71–S102, 10 1990. The Journal of Political Economy, Vol. 98, No. 5, Part 2: The Problem of Development: A Conference of the Institute for the Study of Free Enterprise Systems. (Oct., 1990), pp. S71-S102.

Paul Romer. New Goods, Old Theory, and the Welfare Costs of Trade Restrictions. *Journal of Development Economics*, 43(1):5–38, 1994.

Raaj Sah and Joseph Stiglitz. The Invariance of Market Innovation to the Number of Firms. *The RAND Journal of Economics*, 18(1):98–108, 1987.

Stephen W Salant. Preemptive Patenting and the Persistence of Monopoly: Comment. *The American Economic Review*, 74(1):247–250, March 1984. ISSN 00028282.

Mark Schankerman. How Valuable is Patent Protection? Estimates by Technology Field. *The RAND Journal of Economics*, 29(1):77–107, 1998. ISSN 07416261.

Mark Schankerman and Ariel Pakes. Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period. *The Economic Journal*, 96(384): 1052–1076, December 1986. ISSN 00130133.

Frederic Scherer. Nordhaus' Theory of Optimal Patent Life: A Geometric Reinterpretation. *American Economic Review*, 62(3):422–427, 1972.

Frederic M. Scherer. Research and Development Resource Allocation Under Rivalry. *The Quarterly Journal of Economics*, 81(2):349–394, 1967.

J. Schumpeter. *Captialism, Socialism and Democracy (2nd ed.)*. London Allen and Unwin, 1947.

David Schwartzman. The Burden of Monopoly. *The Journal of Political Economy*, 68 (6):627–630, December 1960. ISSN 00223808.

Suzanne Scotchmer. Standing on the Shoulders of Giants: Cumulative Research and the Patent Law. *The Journal of Economic Perspectives*, 5(1):29–41, 1991. ISSN 08953309.

Suzanne Scotchmer. Protecting Early Innovators: Should Second-Generation Products be Patentable? *The RAND Journal of Economics*, 27(2):322–331, 1996. ISSN 07416261.

Suzanne Scotchmer. On the Optimality of the Patent Renewal System. *The RAND Journal of Economics*, 30(2):181–196, 1999. ISSN 07416261.

Suzanne Scotchmer. *Innovation and Incentives*. MIT, 2004.

Suzanne Scotchmer and Jerry Green. Novelty and Disclosure in Patent Law. *The RAND Journal of Economics*, 21(1):131–146, 1990. ISSN 07416261.

Paul S Segerstrom. Innovation, Imitation, and Economic Growth. *The Journal of Political Economy*, 99(4):807–827, August 1991. ISSN 00223808.

Avner Shaked and John Sutton. Natural Oligopolies. *Econometrica*, 51(5):1469–1483, September 1983. ISSN 00129682.

Carl Shapiro. Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting. In *Innovation Policy and the Economy, Vol. 1*, pages 119–50. MIT Press, 2001.

Steven Shavell and Tanguy van Ypersele. Rewards versus Intellectual Property Rights. *Journal of Law & Economics*, 44(2):525–47, October 2001.

R. David Simpson and Nicholas S Vonortas. Cournot Equilibrium with Imperfectly Appropriable R&D. *The Journal of Industrial Economics*, 42(1):79–92, March 1994. ISSN 00221821.

Michael Spence. Cost Reduction, Competition, and Industry Performance. *Econometrica*, 52(1):101–122, 1984. ISSN 00129682.

Kotaro Suzumura. Cooperative and Noncooperative R&D in an Oligopoly with Spillovers. *The American Economic Review*, 82(5):1307–1320, December 1992. ISSN 00028282.

Pankaj Tandon. Optimal Patents with Compulsory Licensing. *The Journal of Political Economy*, 90(3):470–486, June 1982. ISSN 00223808.

C. T. Taylor and Z. A. Silberston. *The Economic Impact of the Patent System: A Study of the British Experience*. Cambridge University Press, 1973.

David J. Teece. Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6):285–305, December 1986.

Theon van Dijk. Patent Height and Competition in Product Improvements. *The Journal of Industrial Economics*, 44(2):151–167, June 1996. ISSN 00221821.

John Vickers. The Evolution of Market Structure when There is a Sequence of Innovations. *The Journal of Industrial Economics*, 35(1):1–12, September 1986. ISSN 00221821.

Eric Von Hippel. Open Source Projects as Horizontal Innovation Networks - By and For Users. *SSRN eLibrary*, 2002. doi: 10.2139/ssrn.328900.

Michael Waterson. The Economics of Product Patents. *American Economic Review*, 80(4):860–69, September 1990.

Brian D Wright. The Economics of Invention Incentives: Patents, Prizes, and Research Contracts. *The American Economic Review*, 73(4):691–707, September 1983. ISSN 00028282.

Steffen Ziss. Strategic R & D with Spillovers, Collusion and Welfare. *The Journal of Industrial Economics*, 42(4):375–393, December 1994. ISSN 00221821.

Chapter 3

Cumulative Innovation, Sampling and the Hold-Up Problem

3.1 Introduction

> [The] 90-minute documentary [Wanderlust] ... was also a window into the frustrations of making a clip-intensive film dependent on copyright clearance, which has become hugely expensive in the past decade. Initial quotations for the necessary sequences came to more than $450,000, which would have raised by half the cost of the IFC film. ... "Paramount wanted $20,000 for 119 seconds of Paper Moon", Ms. Sams said. "The studios are so afraid of exploitation that they set boundaries no one will cross. Even after the prices were cut, we were $150,000 in the hole."[1]

Cumulative innovation and creativity, whereby new work build upon old, is a pervasive phenomenon. However, it was not until recently that it received significant attention in the literature. The seminal paper in this regard is that of Green and Scotchmer (1995). They introduced a two-stage innovation model in which the second innovation is enabled by, or builds upon, the first. Their paper primarily concerns itself with how rents are divided between innovators at the two stages, in particular with the extent to which the first innovator is (under-)compensated for her contribution (the option value) to the second innovation. They investigate how different policy levers related to intellectual property rights, in particular breadth[2], could be used to affect the bargaining (or its absence) between different innovators and hence the resulting payoffs.

A central feature of their model, as well as subsequent work that extended it (such as Scotchmer (1996)), was an assumption that knowledge of costs and returns, whether deterministic or stochastic, was shared equally by innovators at different stages (i.e. was common knowledge). With common knowledge all mutually beneficial transactions are concluded, using ex ante licenses where necessary to avoid the possibility of hold-up of second-stage innovators.

This assumption, however, is problematic. If all innovators share the same information why do we need different innovators at first and second stages and why concern ourselves with licenses and bargaining if a single innovator could just as easily do it all? The obvious answer is that this assumption is wrong, something suggested by a cursory observation of reality: many different firms engage in innovation precisely because they have specialized skills and knowledge that make it effective for them rather than

[1] The New York Times, May 28, 2006 *No Free Samples for Documentaries: Seeking Film Clips With the Fair-Use Doctrine.*

[2] A monopoly right (intellectual property right) such as a patent or a copyright confers the right to exclude not simply direct copies but also products that are sufficiently similar. The term lagging/leading breath are often used to denote the space of inferior/superior (respectively) products that are excluded by the patent/copyright (i.e. taken as infringing the monopoly).

another firm to engage in a given area.[3] Thus, in this paper we investigate cumulative innovation under asymmetric information, for example, where a first-stage innovator only has a probabilistic prior over the second-stage innovator's cost/values but the second-stage innovator knows them precisely[4].

Our paper takes as a starting point a 'basic' model very similar to that presented by Bessen (2004).[5] Second-stage firms are of two types (high and low value) with the type unobserved by first-stage innovators. With (strong) IP first-stage firms may require second-stage innovators to pay a royalty while with (weak) IP second-stage firms may produce without having to license from first-stage firms. As first-stage firms do not know the type of given second-stage innovator with (strong) IP there may be 'licensing failure' (that is the royalty may be set above the level that a second-stage firm is willing to pay).[6] Thus, there is a trade-off: with IP more first-stage innovation takes place due to the extra royalty income received by first-stage firms but some second-stage innovation may be lost as a result of 'licensing failure' due to high royalty rates.

Such a trade-off is already familiar in the literature and our main reason for presenting it is to provide a benchmark and basis for the more complex 'sampling' model presented in the second section. 'Sampling' is used here broadly to cover any kind of trialling and experimentation activity that is likely to take place before a license can be agreed. The logic here is that there are transactional costs and complexities involved in negotiating and executing a license that mean that it only takes place once some degree of investigation ('sampling') has taken place.

For example, a first-stage innovation might be a 'tool' which the second-stage innovator wishes to use in some manner but is unsure what the most beneficial use for this

[3]See e.g. Eisenberg and Heller (1998); Hall and Ziedonis (2001); Cockburn (2005).

[4]Of course, for consistency, the collective distribution of the values/costs of all second-stage innovators should correspond to the prior of the first innovator.

[5]We differ from Bessen slightly in that his focus is primarily on whether ex ante or ex post licensing occurs. Central to this analysis is his introduction of ex post royalty shares which are the royalty shares that take place in the absence of licensing. These are determined exogenously – perhaps as a policy variable or determined by invent-around costs and other factors – and Bessen shows that the socially optimal ex post royalty share is less than that obtained in ex ante bargaining (and so all licensing should occur ex post). By contrast, in our model we do not have the concept of an ex post royalty share: either a second-stage innnovator obtains a license or she does not (and so then cannot produce).

[6]We note that Bessen uses the term 'holdup' to denote what we term 'licensing failure'. Since he is considering ex-ante licensing his use of the term 'hold-up' differs somewhat from the traditional usage as there are no sunk relationship-specific investments (a binding contract is possible ex-ante). Rather the 'hold-up' is simply that, just like a monopolist facing heterogeneous consumers, a first-stage innovator is facing a set of second-stage innovators with private and heterogeneous values and so may set a profit-maximizing royalty rate that excludes some second-stage innovators from licensing. Since, the 'sampling' case we discuss below resembles more closely a traditional 'hold-up' situation we prefer to reserve that term for use there and to to use 'licensing failure' for the situation described here.

tool is or how valuable usage of the tools will be (this would be particularly relevant to Biotechnology where the issue of research 'tool' licensing is particularly prominent). 'Sampling' in this case would correspond to experimentation by the second-stage innovator in order to determine the best way to use the tool and/or how valuable such usage is. The more 'sampling' a second-stage innovator does, the more likely the resulting use is a high value one.

Alternatively, one could imagine a first-stage innovation is a basic product that the second-stage innovator wishes to extend. Here again, the second-stage innovator while knowing that she wishes to extend a particular first-stage innovation may not necessarily have a clear idea as to how best to do this (or whether the particular idea she has is actually feasible). 'Sampling' would then indicate the trialling and experimentation necessary to reduce these uncertainties and improve the likelihood the result is a good one.

Real-world examples of such situations can be found in several areas. Consider first the example of documentary film-maker wishing to make a film on a particular topic and requiring clips from a particular source. The film-maker will likely need to have expended significant time experimenting with the source footage and weaving it into their work before arriving at the point of seeking a formal license and the more time spent the better the likely end result. Similarly in music, particularly modern music, re-use, and the associated experimentation and trialling, is ubiquitous. In particular, in dance and hip-hop, the act of 'sampling', whereby a small section of a previous work is directly copied and then repeated or reworked in some manner, is the very basis of the genre and, once again, the more time spent experimenting with a particular source track the better the resulting work.[7]

Meanwhile in software it is common for developers to expend significant time trying out and experimenting with an existing product or 'library' in order to determine whether they can extend it or integrate in the way that would be useful or fits with their existing needs. Again significant effort may need to be expended before a formal license is concluded (for example, the license may depend on the exact intended usage), and the more time a user or developer spends 'sampling' the more likely the resulting application is a good one. The same logic applies to other information products such as databases, as well as to research tools in areas like the life sciences. Here a (potential) user may need to spend significant time exploring the content and features of the

[7]More generally all composers whether classical or modern use previous musical, ideas, motifs, and melodies as parts of new works. See e.g. Malcolm Gladwell, The New Yorker, 2004-11-22, *Something Borrowed: Should a charge of plagiarism ruin your life?*, also http://www.low-life.fsnet.co.uk/copyright/part3.htm#copyrightinfringement for information about sampling in dance and hip-hop music.

product, as well as trialling different ways to use and apply it prior to agreeing a formal license, with the likelihood of a good application increasing in the time spent in this way.

Returning to the model, the key point is that sampling here benefits an innovator by increasing the probability of having a high value innovation but it is costly. As it takes place prior to any kind of royalty negotiation it may lead to hold-up: the hold-up of the sampling effort. As a result the presence of IP rights that require second-stage innovators to license may now have another cost in addition to that from traditional 'licensing failure': fearing high royalty rates second-stage firms will reduce the level of sampling they do and thereby reduce the average quality of second-stage innovations. Because this effect operates across all second-stage innovators its consequences for welfare may be substantially greater than the traditional 'licensing failure' problem (which only affects low value second-stage innovators).

Turning to the comparative statics, we find that, in general, the lower the sampling costs or the larger the differential between high and low value second-stage innovations, the more likely it is that a regime *without* intellectual property rights will be preferable. Thus, in the context of this model, technological change which reduces the cost of encountering and trialling new 'ideas' should imply a reduction in the socially optimal level of intellectual property rights such as patents and copyright.

This approach therefore adds another dimension to the question of how profit is divided between innovators at different stages. Seen in this light, it also has direct analogies with existing results related to the question of whether second-stage innovations should be infringing (I) or non-infringing (NI). For example, Denicolo (2000), who extends Green and Scotchmer's model with patent races at each stage, finds that in some circumstances it will be better to make second-stage innovations non-infringing (in this model one trades off faster second-stage innovation with non-infringement against faster first-stage innovation when there is infringement).

It also has a close connection to the recent work of Bessen and Maskin (2006). Similar to this paper they investigate the welfare impact of 'licensing failure' due to asymmetric information in a model of cumulative innovation. Similar to us they show that, with cumulative innovation, in contrast to what occurs in a 'one-shot' model, IP may, in some circumstances, reduce rather than increase innovation (and social welfare). However their focus is rather different from ours (complementarities in research rather than sampling) and their results arise for different reasons. Specifically, in their model there are multiple stages with (the same) two firms at each stage. Each may choose to participate or not in researching the current innovation and the next innovation stage is reached if, and only if, research at the current stage is successful, with success

an increasing function of the number of participating firms. As a result their is an 'externality' from participation in a given stage: though the value of success at the current stage accrues only to the winning firm by enabling subsequent stages (some of which may be won by the other firm) success also increases the other firms expected revenue. As a result, when one firm is excluded from subsequent stages due to 'licensing failure' under an IP regime the effect on welfare can be far more severe than in the one-stage case.

Likewise the present paper also has a connection to the recent paper of Polanski (2007). That presents a 'centipede-type' model of k-stage cumulative innovation and compares 'Open-Source' (OS) and 'Proprietary' (PR) production. The key assumptions there are that (a) producers derive some direct benefit from product improvements independent of any sales income (without this 'Open Source' would never work), and (b) there is only ex-post bargaining between stage producers in the 'Proprietary' mode which generates 'hold-up' problem effects – at each stage a given producer has sunk her costs before bargaining with the next stage producer begins and this problem can 'cumulate' over the innovation chain. Together these generate the main result that either mode of production, in the right circumstances, can be dominant – in the sense of permitting production when the other does not (though obviously for differing parameter values). Again, while some of the results in this paper have a similar flavour to those in Polanski (2007), our model differs substantially in way they arise. Specifically here there are only two stages and the main 'hold-up' issue arises in relation to the second-stage innovators 'sampling' effort and its interaction with licensing failure in the presence of imperfect information about the types of second-stage innovators.

Finally, we should point out that our results are of relevance to a variety of recent policy debates. For example, in December 2006 the Gowers Review of Intellectual Property which had been setup by the UK government to examine the UK's current IP regime, provided, as one its recommendations (no. 11), that "Directive 2001/29/EC [the EU Copyright 'InfoSoc' Directive] be amended to allow for an exception for creative, transformative or derivative works, within the parameters of the Berne Three Step Test." Such a 'transformative use exception' would correspond very closely to the weak/no IP regime considered in the model presented here. Meanwhile in 2005 in the United States, the Supreme Court in *Merck KGaA v. Integra Life Sciences I, Ltd*[8] created a very broad research exemption in relation to pre-clinical R&D. Such a change again corresponds closely in the model to a move towards a weak/no IP regime in which a second-stage product would not infringe on a first-stage firm's patent.

[8]The full opinion is available at http://www.supremecourtus.gov/opinions/04slipopinion.html.

3.2 A Basic Model of Two-Stage Cumulative Innovation

3.2.1 The Model

We adopt a simple model of two stage innovation in which the second innovation builds upon the first in some manner – either as an application or as an extension of it. All agents are risk-neutral and act to maximize profits.

Innovations are described by their net value v (revenue minus costs). Because our interest lies in examining the trade-off between innovation at different stages we make no distinction between social and private value (i.e. there are no deadweight losses) and v may be taken to be both.

We assume the base (first) innovation takes two values: low (v_1^L) and high (v_1^H) with probability $p, (1-p)$ respectively. We assume that $v_1^L < 0$ so that without some additional source of revenue, for example from licensing (see below), the innovation will not be produced. High value innovations have positive stand-alone value, $v_1^H > 0$, and so do not require an outside source of revenue in order to be profitable.

Second-stage innovations also take two values: low (v_2^L) and high (v_2^H) with probability $q, (1-q)$ respectively and $v_2^H > v_2^L > 0$. While the value of a second-stage innovation is known to the innovator who produces it, the value is **not** known to the owner of the first-stage innovation which it builds upon (this could occur because of imperfect information regarding revenue, costs or both). Without loss of generality we shall assume that the number (or measure) of second-stage innovations per first-stage innovation is one (having N second-stage innovations per first-stage innovation would just require replacing v_2^H with Nv_2^H and v_2^L with Nv_2^L). We also assume that $v_1^L + v_2^L \geq 0$ – this ensures that whatever the value of q the overall value generated by a first-stage innovation is positive (the overall value is the stand-alone plus the value of dependent second-stage innovations).[9]

Intellectual Property Rights and Licensing

We wish to consider two regimes: one with (strong) intellectual property rights (IP) and one with weak, or no intellectual property rights (NIP). With intellectual property rights every second-stage innovator will require a license from the relevant first-stage innovator in order to market her product, while without intellectual property rights

[9] Allowing values of v_1^L less than v_2^L does not alter the analysis in any significant way but brings extra complexity to the statement and proof of propositions.

she may market freely without payment or licence.[10]

We assume that the direct returns to the first innovator (v_1) are unaffected by the intellectual property rights regime. This assumption is not as strong as it first appears since simple business stealing, in which the total combined rents of the two stages remain unchanged, could be incorporated into this model simply by increasing p, the proportion of first stage innovations that are low value.[11] Of course, if there is rent dissipation, due, say to further product market competition, this would not be the case and a richer model would be required. Given our need to keep the analysis tractable, and that the focus in this paper is on the division of rents between first and second-stage innovators, we do not take this approach, though we do return to the matter briefly in the conclusion.

Finally, we take the licence to define a lump-sum royalty payment r. This assumption is without loss of generality since, in this model, an innovation is entirely defined by its net value v and there are no other attributes available to use in designing a mechanism to discriminate between types of second-stage innovator.[12] The royalty is set ex-ante, that is prior to the second-stage innovator's decision to invest, and is in the form of a take-it-or-leave it offer by the first-stage innovator.

Sequence of Actions

The sequence of actions in the model is:

1. Nature determines the value type of the first-stage innovator.

2. A first-stage innovator decide whether to invest. If the first-stage innovator does not invest the game ends and all payoffs are zero. Assuming the first-stage innovator invests the game continues.

3. The first-stage innovator sets the royalty rate r (under the no/weak IP regime second-stage innovations do not infringe and so the de facto royalty rate is 0).

4. Nature determines the value type of a second-stage innovator.

[10]Given that we are dealing with cumulative innovation some readers might prefer the infringing (I) vs. non-infringing (NI) dichotomy with its implication of a distinction between 'horizontal' imitation and 'vertical' improvement of a product.

[11]The assumption would also be valid in the case where there is little substitution between the first and second-stage innovation. For example, where the first innovation is a tool used in developing the second-stage innovation.

[12]For example, there are no quantities on which to base a non-linear pricing scheme (fixed fee plus per unit fee royalty). For the same reason there is no opportunity to use type-contingent menus, or any other form of more complex licensing agreement, to increase total royalty income by discriminating between high and low value innovators.

Player	Second-Stage Innovator					
First Stage Innovator	Value Type		Low (q)		High (1-q)	
		Action	NI	I	NI	I
	Low (p)	r	v_1^L , 0	$v_1^L + r$, $v_2^L - r$	v_1^L , 0	$v_1^L + r$, $v_2^H - r$
	High (1-p)		v_1^H , 0	$v_1^H + r$, $v_2^L - r$	v_1^H , 0	$v_1^H + r$, $v_2^H - r$

Table 3.1: Action and Payoff Matrix Assuming First-Stage Innovator Invests. (I/NI = Invest/Do Not Invest, r = Royalty Rate)

5. Given this royalty rate second-stage firms decide whether to invest.

6. Payoffs are realized.

The action/payoff matrix is summarized in Table 3.1.

3.2.2 Solving the Model

Define a constant, α, as follows:

$$\alpha \equiv \frac{v_2^H - v_2^L}{v_2^H}$$

Proposition 3.2.1. *With intellectual property rights, the game defined above has the following Subgame Perfect Nash equilibria. A second-stage innovator invests if and only if its realized value is greater than or equal to the royalty rate (i.e. net profits are non-negative). A first-stage innovator invests and sets a low royalty rate (RL), $r_L = v_2^L$ if the probability of a low value innovation (q) is greater than α and a high royalty rate (RH) $r_H = v_2^H$ if $q \leq \alpha$. When $q = \alpha$ the first-stage innovator may set any royalty of the form r_L with probability x and r_H with probability $1-x$, $x \in [0, 1]$. Thus, there always exist a pure strategy equilibrium and except when $q = \alpha$, this equilibrium is unique.*

Proof. See appendix. □

Proposition 3.2.2. *Without intellectual property rights the game above has the following solution: both types of second-stage innovators invest but, of first-stage innovators, only those that have 'high-value' innovations invest (there are $1 - p$ of these type).*

Proof. Trivial. □

3.2.3 Welfare

To determine welfare we need to know the 'trade-off' between first and second-stage innovations that occurs when revenue is allocated from one to the other by licensing.

	RL	RH
IP	$v_1 + v_2$	$v_1 + (1-q)(v_2^H)$
NIP	$(1-p)(v_1^H + v_2)$	$(1-p)(v_1^H + v_2)$
IP - NIP	$p(v_1^L + v_2^L) + p(v_2 - r_L) \geq 0$	$p(v_1^L + (1-q)v_2^H) - (1-p)qv_2^L$

Table 3.2: Welfare in the Basic Model

As stated above, without royalty income from second-stage innovations a proportion p of first-stage innovations are not produced with average (stand-alone) value v_1^L. The remaining innovations $(1-p)$ are produced irrespective of whether royalty revenue is received and have average value v_1^H.

Let us now consider social welfare in the four possible situations given by (IP, RL), (IP, RH), (NIP, RL), (NIP,RH) as well as the difference in welfare between an intellectual property regime and a no intellectual property regime (IP-NIP). Due to our earlier assumption welfare is determined by calculating total net value. Define for convenience $v_1 = pv_1^L + (1-p)v_1^H$, the average first-stage innovator value (if all innovate), and $v_2 = qv_2^L + (1-q)v_2^H$, the average second-stage innovator value (if all innovate). We summarize the welfare situation in Table 3.2.

3.2.4 Policy Implications

Proposition 3.2.3. *When a low royalty will be set ($q \geq \alpha$) an IP regime is optimal.*

Proof. In the low royalty (RL) situation all second-stage innovations will be produced whether there is IP or not. In that case one wishes to maximize returns to the first innovator and patents do this by transferring rents via licensing. Formally in the low royalty case the welfare difference between patents and no patents (IP-NIP) is:

$$p(v_1^L + r_L) + p(v_2 - r_L)$$

Both of the terms in brackets are positive implying that the intellectual property regime delivers higher welfare than the no intellectual property (NIP) regime. □

The situation when the high royalty will be set is less clear. First, define β as the proportion of the royalty payment to a *low-value* first-stage innovator that would be 'used up' in paying their extra costs:

$$\beta \equiv \frac{-v_1^L}{(1-q)r_H}$$

Note that v_1^L is negative and must be less in absolute terms than the royalty received

$(1-q)r_H$ as we are assuming that the royalty enables low value first-stage innovators to produce. Under this definition $\beta = 1$ corresponds to the case where all of the royalty paid to a low-cost first-stage innovator being used to pay their 'extra' costs while $\beta \approx 0$ means all of the royalty payment is being retained as extra profits (and welfare).

Proposition 3.2.4. *When a high royalty will be set ($q < \alpha$) an intellectual property regime will be preferable to a no intellectual property (NIP) regime if and only if (NB: in fact with equality one would be indifferent):*

$$p \geq \frac{qv_2^L + qv_2^L}{(1-\beta)(1-q)v_2^H} \tag{3.1}$$

$$= \frac{\textit{Licensing Failure Cost}}{\textit{Licensing Failure Cost} + \textit{Surplus From Extra 1st Stage}} \tag{3.2}$$

Proof. From Table 3.2 an IP regime yields higher welfare than an NIP regime if and only if:

$$p(v_1^L + (1-q)v_2^H) \geq (1-p)qv_2^L$$

Making p the subject of this inequality and using β we obtain the stated result. □

We represent the import of these propositions graphically in Figure 3.1, a diagram which shows optimal policy regions as a function of the exogenous probabilities of low value first-stage (p) and second-stage (q) innovations.

Remarks: in the high royalty case (RH) q is the proportion of second-stage innovations that do **not** occur **with** intellectual property rights (due to high royalties and the resulting licensing failure) while p is the proportion of first-stage innovations that do **not** occur **without** intellectual property rights. As first-stage innovations enable second-stage ones when we lose a first-stage innovation we lose all dependent second-stage ones as well. Due to this, when β is low for no intellectual property rights to be preferable q must be substantially higher than p. It is only then that the cost of intellectual property rights, in terms of lost second-stage innovations, will outweigh the gains in terms of more first-stage (and dependent second-stage) innovations.

As β increases the area in which no intellectual property rights are preferable will increase, with the line separating the two regions moving upwards. In the limit as β tends to 1 – which corresponds to all royalty income being used by a low value first-stage innovator to pay costs – the marginal p tends to 1, that is, it is optimal to have intellectual property rights only if all first-stage innovations are of a low value type.

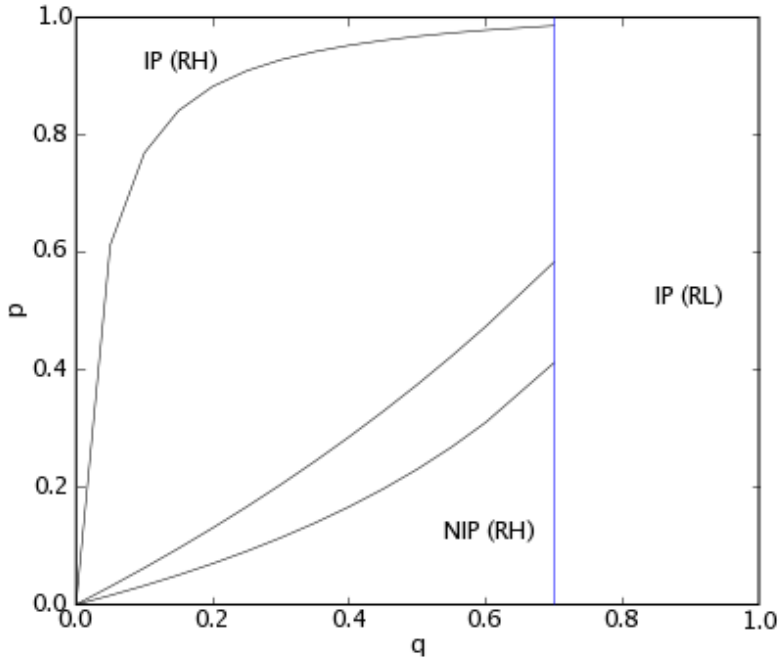

Figure 3.1: Optimal policy as a function of the probabilities of low value first-stage (p) and second-stage (q) innovations. α equals 0.7 so to the right of the line $q = 0.7$ a low royalty will be set (RL) and an IP regime is optimal. To the left of that line we have shown three different 'horizontal' lines which demarcate the boundary between IP being optimal (above the line) and no IP being optimal (below the line). The horizontal lines correspond (in ascending order) to β (the proportion of royalties used up by first stage-innovators) equal to 0, 0.5, and 0.99.

3.3 A Model of Cumulative Innovation with Sampling

3.3.1 The Model

The 'sampling' model differs from the 'basic' model presented in the previous section only in the addition of a single extra period in which sampling by second-stage firms takes place prior to any royalty setting. Formally, we have the following modified sequence of actions (modifications are bolded for clarity):

1. Nature determines the value type of the first-stage innovator.
2. A first-stage innovator decide whether to invest. If the first-stage innovator does not invest the game ends and all payoffs are zero. Assuming the first-stage innovator invests the game continues.
3. **Second-stage innovators chooses their level of sampling k.** (One could think of this, for example, as the number of first-stage products a second-stage firms chooses to investigate via purchase, observation etc).
    - Sampling has constant marginal cost τ.
    - Knowledge of the sampling level chosen by a second-stage firm. There are two possibilities regarding the knowledge of the sampling level available to first-stage innovators. In the first case the first-stage innovator does observe the sampling level. In the second case the first-stage innovator does not observe the sampling level. In what follows we focus on the case where the sampling level is unobserved as we feel this is more realistic though the results are unchanged (and simpler to derive) when it is observed.
4. The first-stage innovator sets the royalty rate r (under the no/weak IP regime second-stage innovations do not infringe and so the de facto royalty rate is 0).
5. Nature determines the value type of a second-stage innovator. As before there are two types of stage 2 firms, high and low value: v_2^H, v_2^L. However, here:
    - **The probability, q, that a second-stage firm is low value is a function of the sampling level: $q \equiv q(k)$.**
    - Properties of $q(k)$: $q' \leq 0$ (otherwise there is no benefit from sampling). There are diminishing returns to sampling: $q'' \geq 0$ and if no sampling takes place all firms are of low value type ($q(0) = 1$). The functional form $q(k)$ is assumed to be common knowledge.

Player			Second-Stage Innovator			
			Sample (k)			
First-Stage Innovator	Value Type		Low (q(k))		High (1-q(k))	
		Action	NI	I	NI	I
	Low (p)	r	v_1^L , $-k\tau$	$v_1^L + r$, $v_2^L - r - k\tau$	v_1^L , $-k\tau$	$v_1^L + r$, $v_2^H - r -$
	High (1-p)		v_1^H , $-k\tau$	$v_1^H + r$, $v_2^L - r - k\tau$	v_1^H , $-k\tau$	$v_1^H + r$, $v_2^H - r -$

Table 3.3: Action and Payoff Matrix Assuming First-Stage Innovator Invests (I/NI = Invest/Do Not Invest, r = Royalty Rate)

6. Given this royalty rate second-stage firms decide whether to invest.

7. Payoffs are realized.

The new action/payoff matrix is shown in Table 3.3.

3.3.2 Solving the Model

Define, as in the basic model, a high royalty to be equal to the value of a high-value second-stage innovation: $r_H = v_2^H$, and a low royalty to be equal to the value of a low-value second-stage innovation: $r_L = v_2^L$.

We begin with a set of preliminary propositions which detail the players best responses before moving on to characterise the equilibrium under both (strong) IP and weak/no IP (NIP).

Proposition 3.3.1 (Second-stage innovator's investment strategies)**.** *A second-stage innovator with value v_X facing a royalty of r will invest if and only if $v_X \geq r$.*

Proof. Just as in the original model second-stage innovator's move with full knowledge of all variables. In this case an innovator of type X invests if and only if net profits from investing, $v_X - r - k\tau$ are greater than $-k\tau$ the payoff from not investing (sampling costs are sunk). Hence the investment strategies are the same as in the basic model: a second-stage innovator invests if and only if $v_X \geq r$. □

Proposition 3.3.2 (First-stage Best-Response Royalty)**.** *Under the IP regime, a first-stage innovator, whose belief about the sampling level is given by the cdf $F(k)$ and where $\bar{q} = \mathbb{E}_F(q(k))$, will set a royalty of the form:*

$$r(k) = \begin{cases} r_L = v_2^L, & \bar{q} > \alpha \\ r_H = v_2^H, & \bar{q} < \alpha \\ \textit{mixed strategy } (r_H, r_L) \textit{ with prob } (x, 1-x), x \in [0,1], & \bar{q} = \alpha \end{cases}$$

where α is as in the basic model, that is the probability such that a first-stage firm is indifferent between setting a high and a low royalty rate:

$$\alpha \equiv \frac{v_2^H - v_2^L}{v_2^H}$$

Proof. See appendix. □

Remark 3.3.3 (Definition of k_α). If a first-stage innovator believes second-stage innovators all play the same pure strategy, k, then we can replace the conditions of the form $\bar{q} <, =, > \alpha$ with the condition that $k >, =, < k_\alpha$ (note the inversion of ordering), where the constant k_α, is the sampling level such that $q(k_\alpha) = \alpha$.

Proposition 3.3.4 (Second-Stage Sampling Level)**.** *Under an IP regime the second-stage innovators best response to a royalty of r, including 'composite' royalties of the form $r = xv_2^H + (1-x)v_2^L, x \in [0,1]$ (that is mixed royalty with r_H played with probability x), is as follows:*

$$k = \begin{cases} k_2, & r \leq r_L = v_2^L \\ k_r, & r_L < r < r_H \\ 0, & r \geq r_H = v_2^H \end{cases}$$

where k_r is defined implicitly by:[13]

$$q'(k_r) = \frac{-\tau}{v_2^H - r}$$

And k_2 is given as follows:[14]

$$k_2 = k_{r_L} = k_{v_2^L} \Rightarrow q'(k_2) = \frac{-\tau}{v_2^H - v_2^L}$$

Proof. See appendix. □

Theorem 3.3.5. *With intellectual property rights (IP) the perfect Bayesian equilibrium of the game defined above falls into one of two cases:*

(i) ***Low royalty case*** *($k_2 \leq k_\alpha$)*

1. *First-stage innovators: both high and low value types invest, believe that second-stage innovators sample at level k_2 and set a low royalty rate.*

[13] If $q'(0) > -\infty$ then for values of r sufficiently close to $r_H = v_2^H$ this equation will have no solution. In such cases define $k_r = 0$.

[14] We use the subscript 2 because this is the level of sampling undertaken in the case where both types of second-stage innovators find it worthwhile to invest.

2. *Second-stage innovators: sample at level k_2 and both high and low value types invest.*

(ii) ***Mixed royalty case*** *($k_2 > k_\alpha$)*

1. *First-stage innovators: both high and low value types invest, believe that second-stage innovators sample at level k_α and set a mixed royalty rate consisting of a high royalty (r_H) with probability x_α and a low royalty (r_L) with probability $(1 - x_\alpha)$ where:*[15]

$$x_\alpha = 1 - \frac{\tau}{-q'(k_\alpha)(v_2^H - v_2^L)}$$

2. *Second-stage innovators: sample at level k_α and invest if and only if the realized value of their innovation is greater than the royalty rate (though the first-stage innovator is playing a mixed strategy the second-stage innovator knows the royalty rate with certainty at the point of investment).*

Proof. See appendix. □

Proposition 3.3.6 (Equilibrium under weak/no IP)**.** *Under weak/no IP the 'sampling' model has the following solution: second-stage innovators sample at level k_2 and both types of second-stage innovators invest. Of first-stage innovators, those that have 'high-value' innovations invest (there are $1 - p$ of these type) and those with 'low-value' innovations do not.*

Proof. Trivial. (Second-stage sampling best-response correspondences have already been derived in Proposition 3.3.4). □

Remark 3.3.7. Recall that k_2 is the sampling level undertaken by a second-stage firm in the case when both high and low value second-stage innovators invest (so it occurs either in the case where there is no IP or when the royalty is sufficiently low). It is also, therefore, the sampling level which maximizes expected second-stage innovation value and, for that reason, the socially optimal sampling level.

3.3.3 Welfare

For the welfare calculations we proceed as in the original model. A proportion p of first-stage innovations are low value ($v_1^L < 0$) and only occur when there is royalty income. Analogously to the basic model define $v_1 = pv_1^L + (1 - p)v_1^H$ and $v_2(k) =$

[15]Note examining the definition of k_2 shows that $k_\alpha < k_2$ guarantees that x_α is non-negative.

$-k\tau+(1-q(k))v_2^H+q(k)v_2^L$ (the expected value generated by a second-stage innovator sampling at level k).

Proposition 3.3.8. *[The Optimal Regime in the Low Royalty Case] In the low royalty case ($k_2 < k_\alpha$) it is optimal to have an IP regime (compared to weak/no IP one). Specifically if the proportion (p) of first-stage innovation that is lost without IP is positive then welfare is higher with IP (otherwise $p = 0$ and both regimes generate the same level of welfare).*

Proof. See appendix. □

This result has a simple intuition behind it. The low royalty case encompasses the situation where the sampling level is fairly low even when the royalty rate faced by second-stage firms is small ($k_2 \leq k_\alpha$) – this may occur because sampling is costly (τ is high) or generates little benefit (v_2^H and v_2^L are close). As a result most second-stage innovations are low value and so a first-stage innovator sets a low royalty rate (r_L). Hence (a) there is no 'licensing failure' and (b) all second-stage firms sample at the optimum rate (k_2). Taken together these mean that, just as with the low royalty case of the simpler model, there are no costs to having strong IP. Since, thanks to the licensing income, there is more (by an amount p) first stage innovation under strong IP than under weak/no IP the strong IP regime is clearly better.

Proposition 3.3.9. *[The Optimal Regime in the Mixed Royalty Case] In the mixed royalty case ($k_2 \geq k_\alpha$) it is optimal to have an IP regime rather than a weak/no (NIP) regime if the proportion (p) of first-stage innovation that does* not *occur under no/weak IP is sufficiently high, specifically:*

$$p \geq p^m \equiv \frac{(v_2(k_2)-v_2(k_\alpha))+x_\alpha q(k_\alpha)v_2^L}{(v_2(k_2)-v_2(k_\alpha))+x_\alpha q(k_\alpha)v_2^L+(v_2(k_\alpha)-x_\alpha q(k_\alpha)v_2^L-(-v_1^L))} \tag{3.3}$$

$$= \frac{\textit{Reduced Sampling Cost} + \textit{Licensing Failure Cost}}{\textit{Reduced Sampling Cost} + \textit{Licensing Failure Cost} + \textit{Surplus from Extra 1st Stage}} \tag{3.4}$$

Proof. See appendix. □

Remark 3.3.10. Reduced Sampling Cost: $v_2(k_2)$ is the average value of second-stage innovations when second-stage firms sample at the unrestricted (and optimal) level k_2. Under the IP regime second-stage firms only sample at level k_α) because of the higher (average) royalty. Thus, the average value of a second-stage innovation is less under the IP regime compared to the weak/no IP regime due to this reduced sampling precisely by the amount: $v_2(k_2) - v_2(k_\alpha)$ (NB: obviously this only applies to those

second-stage innovations associated with the $(1-p)$ first-stage innovations which are produced under both the IP and the weak/no IP regime.)

Licensing Failure Cost: licensing failure occurs when a second-stage firm with a low-value innovation is faced with a high royalty rate. Under the IP regime x_α is the probability that a high royalty is set by a first-stage innovator $q(k_\alpha)$ is the probability a second-stage firm has a low-value innovation. Thus $x_\alpha q(k_\alpha)$ is the probability that licensing failure occurs and when it does the loss equals the potential value of the second-stage innovation: v_2^L.

Surplus from Extra First-Stage Innovation: the plus side of the IP regime is the extra first (and dependent) second-stage innovation that happens because first-stage innovators receive higher incomes. There are a proportion p of low (standalone) value first-stage innovators, who will only invest under the (strong) IP regime. For each such innovation the net surplus generated equals the surplus generated by the second-stage firms plus the net (stand-alone) surplus of a first-stage firm. The expected second-stage surplus equals the average value if all second-stage firms produced (when sampling at k_α: $v_2(k_\alpha)$, minus the surplus of those second-stage firms who are held-up: $x_\alpha q(k_\alpha) v_2^L$. Finally the net standalone surplus of a first-stage firm is $v_1^L < 0$.

Finally, compare equation (3.3) with equation (3.1) from the basic model. The main, and most obvious, difference is that, as well as the standard 'licensing failure cost' of (strong) IP, there is another, additional, cost in the form 'reduced sampling' (and reduced average value of second-stage innovations).

Corollary 3.3.11. *Extending $p^m = 0$ to the low royalty case ($k_2 \leq k_\alpha$) by defining $p^m = 0$ if $k_2 \leq k_\alpha$, we have that an IP regime is optimal if $p > p^m$ and a weak/no IP is optimal if $p < p^m$.*

3.3.4 Policy Implications

Since we do not have any precise estimates for the exogenous parameters such as the sampling cost (τ) or the values of second-stage innovations (v_2^H etc) we cannot make direct statements about which regime would yield higher welfare for a given industry. Instead our approach has been to to pick a 'dependent' variable to focus on (in our case p, the proportion of first-stage innovation 'lost' under weak/no IP) and then derive the 'break-even' or marginal p^m such that if $p = p^m$ we are indifferent in welfare terms between the two regimes.

Our next step is to investigate the comparative statics of the marginal p (p^m) with respect to exogenous variables, in particular the cost of sampling (τ) and the relative value of high (v_2^H) and low type (v_2^L) second-stage innovations.

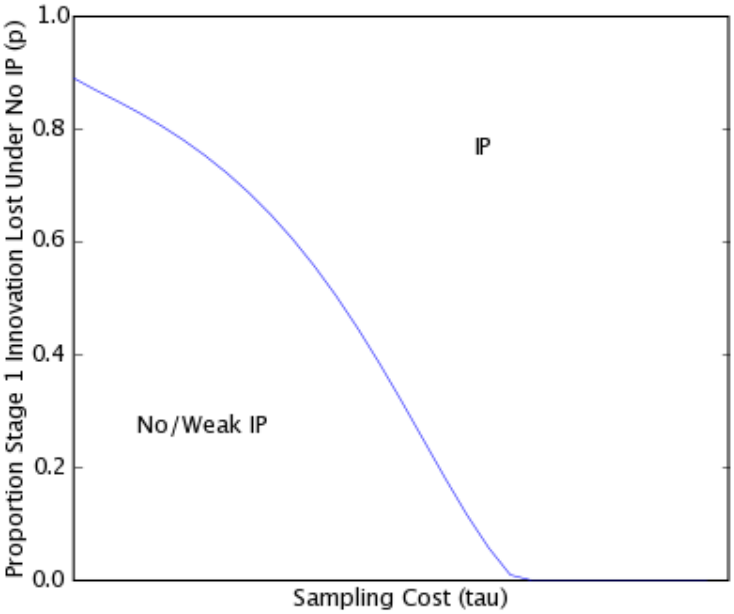

Figure 3.2: Marginal p^m as a function of sampling cost. Ticks on τ axis have been specifically omitted as they would be misleading – any particular value will depend on parametrization, functional form for q etc. However in this specific case we note that $v_2^H = 10, v_2^L = 1.0, q(k) = e^{-k}$ and $p^m = 0$ above 8.0.

Our general results are summarized in Figure 3.2 and Figure 3.3. As we note in the captions one can only indicate the general form as any specific form for p^m will depend on the functional form for q and of course the values of the other exogenous parameters.

Proposition 3.3.12. *The sampling levels k_α and k_2 have the following comparative statics:*

$$\frac{dk_\alpha}{d\tau} = 0 \tag{3.5}$$

$$\frac{dk_\alpha}{dv_2^H} < 0 \tag{3.6}$$

$$\frac{dk_2}{d\tau} < 0 \tag{3.7}$$

$$\frac{dk_2}{dv_2^H} > 0 \tag{3.8}$$

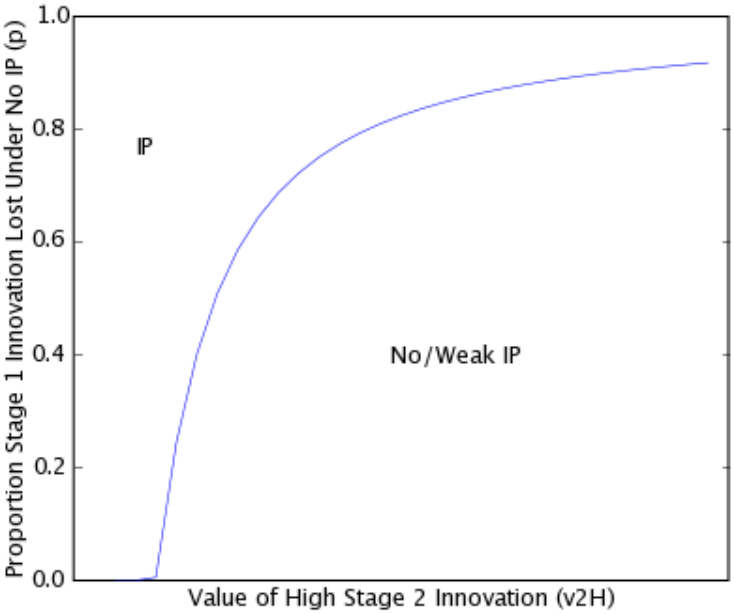

Figure 3.3: Marginal p^m as a function of v_2^H (or equivalently, for fixed v_2^L: $v_2^L - v_2^L$). For the same reasons given in relation to Figure 3.2 ticks on the v_2^H axis have been omitted. However to give the reader some sense of proportion we note that $\tau = 0.5, v_2^L = 1.0, q(k) = e^{-k}$ and $p^m = 0$ below approximately 3.0.

And taking limits:

$$\lim_{\tau \to \infty} k_2 = 0, \lim_{v_2^H \to v_2^{L+}} k_2 = 0, \lim_{\tau \to 0} k_2 = \infty, \lim_{v_2^H \to \infty} k_2 = \infty \tag{3.9}$$

$$\lim_{v_2^H \to \infty} k_\alpha = 0 \tag{3.10}$$

Proof. Recall that we have:

$$q(k_\alpha) = \frac{v_2^H - v_2^L}{v_2^H}$$
$$q'(k_2) = \frac{-\tau}{v_2^H - v_2^L}$$

Given that $q' < 0$ and $q'' > 0$ the results following trivially by simple differentiation. □

Remark 3.3.13. The intuition behind these results is straightforward. k_α is the level of sampling that leaves a first-stage innovator indifferent between charging a high and a low royalty rate. As such it is a function only of the relative values of the two types of innovation (and of q) and does not depend on the cost of sampling at all.

The intuition in the second case is a little more complicated. If we increase v_2^H keeping v_2^L constant we increase the differential between high and low value second-stage innovations. Then the net change in revenue for a first-stage innovator's from switching to a high royalty rate must increase (loss of royalty revenue from low-value second-stage innovations is lower relative to royalty from high-value second-stage innovations). Hence, the proportion of high value second-stage innovations $(1 - q(k))$ at which the switch to a high royalty rate is made is smaller and the corresponding level of sampling (k_α) is smaller.

Coming to k_2, which is the optimal level of sampling (and that performed under a low or zero royalty), we have unsurprisingly that as the cost of sampling goes down the amount of sampling goes up. Similarly, an increase in the relative size of a high value innovation compared to a low value one, increases the benefit of sampling and therefore increases the amount of sampling done.

Combining the differentials with the limits we have that (a) keeping other variables fixed there exists a unique finite τ^* such that for $\tau < \tau^*, k_2 > k_\alpha$ and a mixed royalty is set (conversely for $\tau > \tau^*$ a low royalty is set and $p^m = 0$); (b) similarly there exists a unique v^* such that for $v_2^H > v^*, k_2 > k_\alpha$ and a mixed royalty is set (conversely for $v_2^H < v_2^{H*}$ a low royalty is set and $p^m = 0$). This then demonstrates the validity of the right-hand part of Figure 3.2 and the left-hand part of Figure 3.3 where we have $p^m = 0$.

What occurs then if $k_2 > k_\alpha$ and we are in the mixed royalty case?

Proposition 3.3.14. *Assuming $k_2 > k_\alpha$ (i.e. τ sufficiently small or v_2^H sufficiently large) then:*

$$p \geq p^m \equiv \frac{(v_2(k_2) - v_2(k_\alpha)) + x_\alpha q(k_\alpha) v_2^L}{(v_2(k_2) - v_2(k_\alpha)) + x_\alpha q(k_\alpha) v_2^L + ((v_2(k_\alpha) - x_\alpha q(k_\alpha) v_2^L) - (-v_1^L))}$$

And we have that:

$$\frac{dp^m}{d\tau} < 0$$
$$\frac{dp^m}{dv_2^H} > 0$$

That is, the marginal level of first-stage innovation lost under weak/no IP (that is the level such that above this an IP regime is optimal) is (a) decreasing in sampling costs (b) increasing in the relative size of high value to low value second-stage innovations.

Proof. See appendix. □

Informally this result can be explained as follows. Reductions in sampling costs will increase the 'optimal' level of sampling (k_2) relative to the restricted level of sampling (k_α). This in turn increases the cost of intellectual property rights arising from (a) loss of second-stage innovations due to licensing failure ($x_\alpha \cdot q(k_\alpha)$); (b) lower average value of second-stage innovations ($v_2(k_2) - v_2(k_\alpha)$); while having no effect on the surplus from extra first-stage innovations under IP. As a result the welfare under weak/no IP rises relative to the welfare under IP and the marginal p must rise.

Similarly if the relative size of high value second-stage innovation compared to a low value one rises this (a) increases the 'optimal' level of sampling (k_2) relative to the restricted level of sampling (k_α) (b) directly increases the benefit of sampling. This again increases the sampling cost and the licensing failure cost but reduces the surplus from second-stage innovations under IP. As a result welfare under weak/no IP rises relative to that under IP and the marginal p must rise.

This result then establishes the validity of the rest of Figures 3.2 and 3.3 and implies the following corollaries regarding how the optimal policy regime in relation to intellectual property rights varies in response to changes in the exogenous environment:

Corollary 3.3.15. *Reducing sampling costs make it more likely that a freer (weak/no intellectual property rights) regime will be optimal.*

Proof. Follows from previous propositions as summarised in Figure 3.2. □

Corollary 3.3.16. *Increasing the differential between high and low value second-stage innovations (which could be interpreted as sampling becoming more important for product quality) makes it more likely that a freer (no intellectual property rights) regime will be optimal.*

Proof. Follows from previous propositions as summarised in Figure 3.3. □

Remark 3.3.17. Most studies of the value of intellectual property rights (copyrights or patents) indicate that their distribution is highly skewed with a few very high value works and many low value works. This suggests that $v_2^H >> v_2^L$.

3.4 Conclusion

In this paper we have shown how asymmetric information about the value of follow-on innovations, combined with intellectual property rights such as patents, can result in licensing failure and hold-up. Presenting the policy decision as a choice between having or not having intellectual property rights, we have shown that, in contrast to parts of the previous literature, in some circumstances it may be optimal not to have

intellectual property rights. For whilst intellectual property rights help transfer income from second-stage to first-stage innovators they can also lead to licensing failure and hold-up with a resulting reduction in second-stage innovation.

In the first, and simpler, model presented, the basic results were summarized in Figure 3.1, which plotted optimal policy as a function of the exogenous variables (the probabilities of high or low value innovations occurring at the two different innovation stages). Intellectual property rights in this model had two contrasting effects. On the one hand, there are the benefits of increased first-stage innovation as revenue is transferred to first-stage innovators from second-stage ones. On the other hand, there are costs in terms of fewer second-stage innovations due to licensing failure. In some circumstances the benefits will exceed the costs and we should have (stronger) intellectual property rights. In other cases, they will not and we should have weaker (or no) intellectual property rights. In particular, we showed that, if the probability of a low value second-stage innovation was high enough (but not too high), compared to the probability of a low value first-stage innovation, then a regime without intellectual property rights would be preferable.

Next, we extended this basic model by introducing 'sampling'. We demonstrated the existence of a perfect Bayesian equilibrium and showed that (strong) IP may restrict the level of sampling below what would be socially optimal. Therefore, in addition to the basic trade-off mentioned above between more first-stage innovations and fewer second-stage ones, there is the additional factor: those second-stage innovations which occur have lower average value due to a lower level of sampling. Examining this trade-off, we find that the lower the cost of sampling and the greater the differential between the low and high values of second-stage innovations, the more likely it is that a regime *without* intellectual property rights will be preferable.

Thus, technological change which reduces the cost of encountering and trialling new 'ideas' should imply a reduction in the socially optimal level of intellectual property rights such as patents and copyright. A perfect case of such technological change in recent years can be found in the rapid advances in computers and communications. These advances have, for example, dramatically reduced the cost of accessing and re-using cultural material, such as music and film, as well as greatly increasing the number of 'ideas' that a software developer can encounter and trial. Concrete policy actions that could be taken in line with these conclusions include extending 'fair-use' (fair-dealing) provisions in copyright law to increase the degree of reuse that would be permitted without the need to seek permission and excluding software and business methods from patentability.

Finally, we should emphasize that there remains plentiful scope to improve and

extend the present paper. For instance, it was assumed that the non-royalty income for the first-stage and second-stage innovator was unaffected by the intellectual property rights regime.[16] However this is unlikely to be the case and the model could be improved by the inclusion of the direct effect of no (or weaker) intellectual property rights on the revenue of the first-stage (and second-stage) innovator.

It would also be useful to extend the analysis to the case of a continuous distribution of innovation values, as well as to investigate the consequences of making sampling costs a function of the intellectual property rights regime. It would also be valuable to examine what occurs when the structure of innovation is more complex, for example by having second-stage inventions incorporate many first-stage innovations (a componentized model) or having heterogeneity across innovations with some developments used more than others. Finally, one of the most important extensions would be to properly integrate transaction costs into the analysis. Transaction costs relating to both the acquisition of information and the execution of contracts are significant and without them we lack a key element for the furtherance of our understanding of the process of innovation both in this model and in general.

3.A Proofs

3.A.1 Proof of Proposition 3.2.1

Proof. We are considering only subgame perfect nash equilibria so we may begin at the final stage of the game and work backwards. Given a royalty level of r, at the final stage, a second-stage innovator of type X faces a payoff of $v_2^X - r$ if she invests and 0 if she does not. Thus, a second-stage innovator, seeking to maximize profits will invest if and only if $v_2^X \geq r$ (formally, they are indifferent if $r = v_2^X$. However if they do not invest when $v_2^X = 0$ there will be no equilibrium of the overall game).

Given this, by simple dominance and focusing on pure strategies, a first-stage innovator must EITHER (a) set a low royalty rate $r_L = v_2^L$ which will lead to investment by all second-stage innovations; OR (b) set a high royalty rate $r_H = v_2^H$ which will result in investment only by high value second-stage innovations. In the first case the payoff is r_L while in the second it is $(1-q)r_H$. Thus, a low royalty rate should be chosen if and only if (assuming that if payoffs are equal a low royalty is chosen):

$$r_L \geq (1-q)r_H \iff q \geq \frac{r_H - r_L}{r_H} = \alpha$$

[16] As discussed in detail above, while we do allow for business stealing between the first and second-stage innovators we do not allow for general rent dissipation from wider product market competition.

Since any mixed royalty strategy must consist of some combination of r_L and r_H we have immediately that a proper mixed strategy is only possible when $r_L = (1-q)r_H$, that is if $q = \alpha$.

Finally, total royalty income to a first-stage innovator is at least $r_L = v_2^L$. Thus, total net income for a low-value first-stage innovator is at least $v_1^L + r_L = v_1^L + v_2^L > 0$ (by assumption) – and net income for a high-value first-stage innovator is obviously greater. Hence both types of first-stage innovator will invest. □

3.A.2 Proof of Proposition 3.3.2

Proof. Given a first-stage innovator believes $F(k)$, the expected probability that a second-stage firm is low value is $\mathbb{E}_F(q(k)) = \bar{q}$. By subgame perfection a first-stage innovator knows that, once a seond-stage firm discovers its type, its best response to a given royalty will be as stated in Proposition 3.3.1. In particular, if the royalty rate is set to be less than or equal to the second-stage low value (v_2^L) all second-stage innovators will license, if a royalty is above this but less than or equal to the second-stage high value (v_2^H) then only high value firms will license ($1-\bar{q}$ of them) and if the royalty is higher than this no second-stage firms will license. Then, letting $G(r)$ be the cumulative distribution function over royalties representing the first-stage innovator's mixed strategy, the expected payoff to a first-stage innovator is:

$$\Pi_1(G(r)) = \int_0^{v_2^L} r \cdot dG(r) + (1-\bar{q}) \int_{v_2^L}^{v_2^H} r \cdot dG(r) + 0 \cdot \int_{v_2^H}^{\infty} r \cdot dG(r))$$

Maximizing with respect to $G(r)$ immediately gives that, just as for the basic model, an optimal mixed strategy can only consist of some combination of the pure strategy $r_L = v_2^L$ and the pure strategy $r_H = v_2^H$. Let us suppose that these two pure strategies, r_H, r_L, are played with probability $x, 1-x$ respectively. Revenue from royalties is then:

$$r_L(1-x) + (1-\bar{q})r_H x = r_L + x \cdot ((1-\bar{q})r_H - r_L)$$

Maximizing revenue requires $x = 0$ if the term in brackets is less than zero, $x = 1$ if the term in brackets is greater than 0, and allows any value of x if the term in brackets is zero. By the definition of α (see above) these conditions correspond precisely to $\bar{q}$ (the expected probability of a low value innovation) being less than, greater than or equal to α. Hence, the first-stage innovator's royalty response as a function of their belief about the level of sampling is of the form stated. □

	$r \leq r_L$	$r_L < r < r_H$	$r \geq r_H$
$\Pi(k)$	$-k\tau - r + q(k)v_2^L + (1-q(k))v_2^H$	$-k\tau + (1-q(k))(v_2^H - r)$	$-k\tau$

Table 3.4: Payoff for Second Stage Innovator

3.A.3 Proof of Proposition 3.3.4

Proof. Using the optimal investment stage determined in Proposition 3.3.1, for a given sampling level k, payoffs as a function of the royalty levels are as in Table 3.4.

Suppose second-stage innovator plays a strategy given by the cdf $F(k)$ and a first-stage innovator sets a royalty defined by a cumulative distribution function $G(r)$. Then the payoff to a second-stage innovator is as follows (where expectations are taken with respect to F and q is short for $q(k)$):

$$\begin{aligned}\Pi_2(F(k)) &= \mathbb{E}\left(-\tau k + \int_0^{r_L} qv_2^L + (1-q)v_2^H - r dG(r) + \int_{r_L}^{r_H} (1-q)(v_2^H - r)dG(r) + \int_{r_H}^{\infty} 0 dG(r)\right) \\ &= \mathbb{E}\left(-\tau k - q\{G(r_H)v_2^H - G(r_L)v_2^L - \int_{r_L}^{r_H} r dG(r)\} + G(r_H)v_2^H - \int_0^{r_H} r dG(r)\right)\end{aligned}$$

Claim: Second-stage innovators play pure strategies.

Proof: q is convex so $-q$ is concave. Suppose we have a mixed strategy $F(k)$ with $\mathbb{E}(k) = \bar{k}$ then $-\bar{q} = \mathbb{E}(-q(k)) \leq -q(\bar{k})$ with equality if and only if $F(k)$ is a point distribution (i.e. corresponds to a pure strategy). Substituting:

$$\begin{aligned}\Pi_2(F(k)) &= \mathbb{E}_F(-\tau k + G(r_H)v_2^H - q(G(r_H)v_2^H - G(r_L)v_2^L) - \int_0^{r_H} r dG(r) \\ &= -\tau\bar{k} - \bar{q}\cdot(+\text{ve}) + \text{const} \\ &\leq -\tau\bar{k} - q(\bar{k})\cdot(+\text{ve}) + \text{const}\end{aligned}$$

(With equality iff and only if $F(k)$ is a point distribution with $k = \bar{k}$ with probability 1). Thus for any properly mixed strategy $F(k)$ we can always achieve a higher payoff by playing the pure strategy $\bar{k} = \mathbb{E}(k)$. □

Thus, in what follows we may confine our attention to pure strategies k. Returning to the payoff function we first note that if royalty (or royalties in a mixed strategy) are all greater than r_H (formally the support of $G(r)$ lies entirely above r_H) then the optimal sampling level is zero ($\Pi_2(k) = -k\tau$).

When this is not the case we have the first order condition is:[17]

[17]The second order condition, $\Pi'' \leq 0$, is easily checked: $\Pi'' = -q''(k)\cdot(+\text{ve}) < 0$ since, by assumption, $q''(k) > 0$.

$$q'(k) = \frac{-\tau}{G(r_H)v_2^H - G(r_L)v_2^L - \int_{r_L}^{r_H} r dG(r)}$$

For ease of reference define S as the denominator in the previous equation. We shall look at several special cases as follows:

(i) $r \leq r_L$. Then $G(r_H) = G(r_L) = 1$ and we have $S = v_2^H - v_2^L$. The profit-maximizing k therefore equals k_2 where (as defined above):

$$q'(k_2) = \frac{-\tau}{v_2^H - v_2^L}$$

The intuition here is simple: both firms always invest and pay the royalty. Thus, in terms of the payoff sampling will only affect the value type and the sampling level will be chosen so that the marginal gain in terms of lower costs, $q'(k)(v_2^H - v_2^L)$, equals the marginal sampling costs, τ.

(ii) $r_L < r < r_H$. Here $G(r_L) = 0, G(r_H) = 1$ and we have $S = v_2^H - r$ and the optimal $k \equiv k_r$ solves:

$$q'(k_r) = \frac{-\tau}{v_2^H - r}$$

(iii) r_H played with probability x and r_L with probability $(1-x)$. Then $G(r_L) = (1-x), G(r_H) = 1$. Define the 'composite' royalty $r = xr_H + (1-x)r_L = xv_2^H + (1-x)v_2^L$ then we have $S = v_2^H - (1-x)v_2^L - xr_H = (1-x)(v_2^H - v_2^L) = v_2^H - r$. So the optimal sampling level is $k \equiv k_r$ where r is the composite royalty.

□

3.A.4 Proof of Theorem 3.3.5

Proof. We will solve for a subgame perfect Bayesian nash equilibrium by recursing backwards through the game.

In previous propositions we have already derived the best-response correspondences (where the royalty best-response is defined in terms of *beliefs* about sampling rather than the actual sampling level). We have also shown second-stage firms will always play a pure strategy (i.e. choose a single sampling level). Furthermore, at the sampling stage all second-stage firms are the same, hence all second-stage firms will choose the same pure sampling strategy. Thus, a first-stage innovator's beliefs (to be consistent) must be single-valued and we may rewrite the royalty best-response correspondence in

terms of their belief as to the sampling level (k):[18]

$$r(k) = \begin{cases} r_L = v_2^L, & k < k_\alpha \\ r_H = v_2^H, & k > k_\alpha \\ \text{mixed strategy } (r_H, r_L) \text{ with prob } (x, 1-x), x \in [0,1], & k = k_\alpha \end{cases}$$

Case 1: $k_2 \leq k_\alpha$**.** There are three possibilities for the beliefs of a first stage innovator regarding the sampling level of second-stage firms:

(i) $k > k_\alpha$. Hence the first-stage innovator would set a high royalty rate. Then second-stage innovator's best response is $k = 0$ and beliefs will be inconsistent. Thus, there cannot be an equilibrium with such beliefs.

(ii) $k < k_\alpha$. In this case the best response of a first-stage innovator is to set a low royalty (r_L) in which case second-stage firm must choose a sampling level $k = k_2$. Thus, for beliefs to be consistent, a first-stage innovator must believe $k = k_2$ and the equilibrium is as claimed.

(iii) $k = k_\alpha$. In this case a first-stage innovator's best response correspondence consists of all mixed strategies: r_H with probability x, r_L with probability $1 - x$ for $x \in [0, 1]$. Now a second-stage innovator (if behaving optimally) never samples above the level k_2 and will sample strictly below k_2 if the first-stage innovator plays any strategy in which r_H is played with positive probability. Hence if beliefs are to be consistent we must have (a) $k_2 = k_\alpha$ and (b) $x = 0$ (i.e. a low royalty is always set). In such a case the equilibrium is again as claimed.

Case 2: $k_2 > k_\alpha$**.** There are three possibilities for the beliefs of a first stage innovator regarding the sampling level of second-stage firms:

(i) $k > k_\alpha$. Just as in the first case this leads to inconsistent beliefs and so cannot be an equilibrium.

(ii) $k < k_\alpha$. In this case the best response of a first-stage innovator is to set a low royalty (r_L) in which case second-stage firm must choose a sampling level $k = k_2$. But $k_2 > k_\alpha$. Thus, beliefs will be inconsistent and this cannot be an equilibrium.

(iii) $k = k_\alpha$. In this case a first-stage innovator best response correspondence consists of all mixed strategies: r_H with probability x, r_L with probability $1 - x$ for

[18] At the sampling stage all second-stage firms are the same and their best-response correspondence is single-valued. Hence all second-stage firms must have the same sampling strategy and a first-stage innovator's belief

$x \in [0,1]$. Denote the corresponding composite royalty by $r(x) = xr_H + (1-x)r_L$. Then for an equilibrium (with consistent beliefs) we must find an x such that the best-response sampling level equals k_α. Formally, using the notation of Proposition 3.3.4 we must find an x such $k_{r(x)} = k_\alpha$. The best response sampling level is defined implicitly by:

$$q'(k) = \frac{-\tau}{(1-x)(v_2^H - v_2^L)}$$

Since $q' < 0$ we have, denoting $k(x)$ as the implicit solution as a function of x, that $k'(x) < 0$ (intuitively a higher average royalty lowers sampling). Since $k(0) = k_2 > k_\alpha$ and that $k(1) = 0$ (as $x \to 1$ the RHS of the above takes arbitrarily large negative values), by the intermediate value theorem and the monotonicity of $k(x)$, there must exist a unique $x_\alpha \in (0,1)$ such that $k(x_\alpha) = k_\alpha$. Replacing $q'(k)$ by $q'(k_\alpha)$ and rearranging we have as claimed that:

$$x_\alpha = 1 - \frac{\tau}{-q'(k_\alpha)(v_2^H - v_2^L)}$$

First-stage innovators investment strategy: finally as with our basic model first-stage innovators of both types invest because with royalty income net profits will be non-negative. □

3.A.5 Proof of Proposition 3.3.8

Proof. Analogously to the low royalty case in the basic model, in this situation all second-stage innovators invest so (a) there is no licensing failure (b) second-stage firms sample at the optimal level (k_2). At the same time, intellectual property allows some first-stage innovators to engage in production who wouldn't be able to do so otherwise. Hence an IP regime will deliver higher welfare.

Formally, the welfare difference between the IP and NIP regime is net surplus associated with the p extra first-stage innovations that occur under IP:

$$p((v_1^L + r_L) + (v_2(k_2) - r_L))$$

Both the first term (by the assumption that the royalty is sufficient to allow production) and the second (since second-stage innovators are making non-negative profits) are positive. Hence, if $p > 0$ the sum is positive and welfare is higher with intellectual property. □

3.A.6 Proof of Proposition 3.3.9

Proof. In this case comparing the IP to the no/weak IP regime we have the following differences:

(+) Under IP there are (p) extra first-stage (and dependent second-stage) innovation because the royalty income allows some first-stage innovators to produce who would not otherwise:

$$p \underbrace{(v_1^L + v_2(k_\alpha) - x_\alpha q(k_\alpha) v_2^L)}_{\text{surplus per extra first stage innovation}}$$

(-) For the $(1-p)$ first-stage innovations that occur under both IP and no/weak IP there are fewer associated second-stage innovations due to licensing failure (licensing failure cost) and the innovations are of lower average value due to reduced sampling (reduced sampling cost):

$$-(1-p)(\underbrace{(v_2(k_2) - (v_2(k_\alpha))}_{\text{Reduced Sampling Cost}} + \underbrace{x_\alpha q(k_\alpha) v_2^L}_{\text{Licensing Failure Cost}})$$

An IP regime is optimal compared to a weak/no IP (NIP) if the first effect is larger than the second (and vice versa):

$$\begin{aligned} & p(v_1^L + v_2(k_\alpha) - x_\alpha q(k_\alpha) v_2^L) - (1-p)(v_2(k_2) - v_2(k_\alpha) + x_\alpha q(k_\alpha) v_2^L) \geq 0 \\ \Leftrightarrow & p \geq p^m \equiv \frac{(v_2(k_2) - v_2(k_\alpha)) + x_\alpha q(k_\alpha) v_2^L}{(v_2(k_2) - v_2(k_\alpha)) + x_\alpha q(k_\alpha) v_2^L + ((v_2(k_\alpha) - x_\alpha q(k_\alpha) v_2^L) - (-v_1^L))} \end{aligned}$$

Where p^m has been defined as the probability of a low value first-stage innovation which leaves one indifferent between having and not having intellectual property rights. □

3.A.7 Proof of Proposition 3.3.14

Proof. Define:

$$\begin{aligned} S &= \text{Higher Sampling Cost} = (v_2(k_2) - v_2(k_\alpha)) \\ H &= \text{Licensing Failure Cost} = x_\alpha q(k_\alpha) v_2^L \\ E &= \text{Surplus per Extra Stage 1} = v_2(k_\alpha) - x_\alpha q(k_\alpha) v_2^L - (-v_1^L) \end{aligned}$$

Then,

$$p^m = \frac{S+H}{S+H+E}$$

Examining the differentials of S, H, E we have:

$$\begin{aligned}
\frac{dS}{d\tau} &= \frac{\partial}{\partial\tau}(v_2(k_2) - v_2(k_\alpha)) + \frac{dv_2(k_2)}{dk_2}\frac{dk_2}{d\tau} - \frac{dv_2(k_\alpha))}{dk_\alpha}\frac{dk_\alpha}{d\tau} \\
&= (-) + (+\cdot -) + (+\cdot 0) = - \\
\frac{dH}{d\tau} &= \frac{dx_\alpha}{d\tau}(\cdots) + (\cdots)\frac{dk_\alpha}{d\tau} = (-\cdot +) + 0 = - \\
\frac{dE}{d\tau} &= \frac{dE}{dk_\alpha}\frac{dk_\alpha}{d\tau} = (\cdots)\cdot 0 = 0
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{dS}{dv_2^H} &= + \\
\frac{dH}{dv_2^H} &= + \\
\frac{dE}{dv_2^H} &= -
\end{aligned}$$

For the last equation note, that by definition of k_α, $v_2(k_\alpha) = (1 + q(k_\alpha))v_2^L - k_\alpha\tau$ and that for $k < k_2, v'(k) > 0$ so that:

$$\frac{dv_2(k_\alpha)}{dv_2^H} = \frac{\partial v_2(k_\alpha)}{\partial v_2^H} + v'(k_\alpha)\frac{dk_\alpha}{dv_2^H} = 0 + (+\cdot -) = -$$

Putting these derivatives together with the derivative of p^m with respect to S, H, E we have the required result. □

Bibliography

James Bessen. Hold-up and Patent Licensing of Cumulative Innovations with Private Information. *Economics Letters*, 82(3):321–326, 2004.

James Bessen and Eric Maskin. Sequential Innovation, Patents, and Innovation. NajEcon Working Paper Reviews 321307000000000021, www.najecon.org, May 2006.

Howard F Chang. Patent Scope, Antitrust Policy, and Cumulative Innovation. *The RAND Journal of Economics*, 26(1):34–57, 1995. ISSN 07416261.

Iain Cockburn. Blurred Boundaries: Tensions Between Open Scientific Resources and Commercial Exploitation of Knowledge in Biomedical Research, 2005.

Victor Denicolo. Two-Stage Patent Races and Patent Policy. *Rand Journal of Economics*, 31:488–501, 2000.

R. Eisenberg and M. Heller. Can Patents Deter Innovation? The Anticommons in Biomedical Research. *Science*, 280(5364):690–701, 5 1998.

Nancy Gallini. Patent Policy and Costly Imitation. *Rand Journal of Economics*, 23 (1):52–63, 1992.

Jerry Green and Suzanne Scotchmer. On the Division of Profit between Sequential Innovators. *Rand Journal of Economics*, 26(1):20–33, 1995.

Bronwyn Hall and Rosemarie Ziedonis. The patent paradox revisited: an empirical study of patenting in the U.S. semiconductor industry, 1979-1995. *Rand Journal of Economics*, 32(1):101–128, 2001.

Carmen Matutues, Pierre Regibeau, and Katharine Rockett. Optimal Patent Design and the Diffusion of Innovations. *Rand Journal of Economics*, 27(1):60–83, 1996.

Peter Menell and Suzanne Scotchmer. Intellectual Property, 6 2005. forthcoming, Handbook of Law and Economics. Mitch Polinsky and Steven Shavell, eds. Amsterdam: Elsevier.

Ted O'Donoghue, Suzanne Scotchmer, and Jacques Thisse. Patent Breadth, Patent Life, and the Pace of Technological Improvement. *Journal of Economics and Management Strategy*, 7:1–32, 1998.

Arnold Polanski. Is the General Public Licence a Rational Choice? *The Journal of Industrial Economics*, 55(4):691–714, Dec 2007.

Suzanne Scotchmer. Protecting Early Innovators: Should Second-Generation Products be Patentable? *The RAND Journal of Economics*, 27(2):322–331, 1996. ISSN 07416261.

Chapter 4

Innovation and Imitation with and without Intellectual Property Rights

4.1 Introduction

Repeated surveys, such as Levin et al. (1987), Mansfield (1985), Cohen, Nelson, and Walsh (2000), and Arundel (2001), show that firms appropriate returns from innovation using a variety of methods including secrecy, lead time, marketing and sales, learning curve advantages and patents. Furthermore, they also suggest that for most industries (with a few notable exceptions) patent protection is of low importance. As Hall (2003) summarizes (p. 9): 'In both the United States and Europe, firms rate superior sales and service, lead time, and secrecy as far more important than patents in securing the returns to innovation. Patents are usually reported to be important primarily for blocking and defensive purposes.'

Of particular interest is the finding that imitation is a costly process both in terms of time and money, and one, furthermore, upon which the effect of a patent – if it has any effect at all – is to increase its cost not to halt it entirely. Perhaps most striking in this respect are Tables 8 (p. 810) and 9 (p. 811) of Levin et al. (1987) which summarize, respectively, reported cost of imitation (as a percentage of innovator's R&D expenditure) and time to imitate. For example, of the processes surveyed which were not protected by patents fully 88% had an imitation cost which was more than 50% of the innovator's initial outlay. For major products the analogous figure was 86%. Imitation also takes time: 84% of unpatented processes took 1 year of longer to imitate, while for products the analogous figure 82%.[1]

Such results indicate that for many innovations, even without patent protection, imitation involves substantial cost and delay.[2] Given this, as well as the strong impact the assumption of costless imitation has on our conclusions, it would seem important to investigate the consequences of weakening this presumption and, in particular, the possibilities of innovation without intellectual property rights.

However much of the existing theoretical literature has tended to assume 'perfect' nonrivalry, that is, that an innovation (or creative work) once made may be costlessly, and instantaneously, reproduced. The assumption is most often evident in the claim, which follows directly from it, namely that without the provision of intellectual property rights such as patents and copyrights no innovation would be possible.

For example, Nordhaus (1969) (and following him Scherer (1972)), in what is considered to be one of the founding papers of the policy literature, implicitly assume that without a patent an innovator gains no remuneration. Similarly, Klemperer (1990) in

[1] Of course, one must be cautious in interpreting such figures given the likely selection bias in deciding whether to patent or not – it is precisely those innovations which are hard to imitate without a patent which will not be patented.

[2] As emphasized by Dosi (1988) the distinction between innovation and imitation is often highly blurred and that imitation itself is a creative process.

his paper on patent breadth makes clear his assumption of costless imitation[3](p. 117): 'For simplicity, I assume free entry into the industry subject to the noninfringement of the patent and that knowledge of the innovation allows competitors' products to be produced *without* fixed costs and at the same constant marginal costs as the patentholder's product. Without further loss of generality, I assume marginal costs to be zero.' (Emphasis added). Many similar examples can easily be supplied in which imitation without intellectual property rights is implicitly, or explicitly, assumed to be 'trivial'.[4]

This paper, by contrast, provides a simple theoretical model in which costly imitation is central. Combined with first-mover advantage for the innovator we show that a significant amount of innovation takes place in the absence of intellectual property rights – even when imitation is cheaper than innovation. In addition we provide an easy and intuitive way to conceptualize, and model, the overall space of innovations which allows us to compare in a straightforward manner the relative performance of regimes with and without intellectual property rights, both in terms of innovation and welfare. This approach supplies several novel insights.

First, that as innovation costs fall 'allowable' imitation costs (that is imitation costs that still result in innovation being made) fall even faster. Thus, if the cost of innovation (relative to market size) differs between industries, then, even if relative imitation costs are the same, there will be very substantial difference in the impact of intellectual property rights. In particular, in the industry with lower innovation costs the gains for innovation and welfare with intellectual property rights will be much lower (and for welfare could even be negative).[5] As such, a main point of this paper is to show how the impact (and benefits/costs) of intellectual property rights may vary in a systematic way across industries. In particular there will be industries in which intellectual property rights are necessary – and industries where they are not – and this paper presents one basis for a taxonomy to sort out which is which.

Second, and relatedly, comparing regimes without and with intellectual property rights we show that the welfare ratio is systematically higher than the innovation ratio.[6] Moreover, this is not simply for the well-known reason that (conditional on the

[3]Though it should be noted that it is possible to interpret the travel cost incurred by consumers in Klemperer's model as some form of 'design-around' or imitation cost that must be paid by competing firms. Nevertheless, in Klemperer's model, absent IP the innovator's gross profits (excluding the sunk cost of innovation) will be driven to zero by competition. As a result, anticipating a net loss, an innovator would not enter.

[4]See e.g. Scotchmer and Green (1990); Hopenhayn and Mitchell (2001); Menell and Scotchmer (2005).

[5]Consider, for example, pharmaceuticals compared to software. Starting a pharmaceutical (or biotech) company requires very substantial investment on the order of millions of euros while a software startup may need only a few tens of thousands of euros.

[6]The innovation ratio is the innovation level without intellectual property rights versus the level

innovation occurring) without intellectual property rights greater competition results in increased output and lower deadweight losses. Rather, there is an additional factor, namely that the set of innovations occurring under an IP regime are, on average, less socially valuable because they have higher fixed costs of creation. Specifically, the model allows us to clearly distinguish three sources of welfare differences between the two regimes: first, less innovation occurs without intellectual property rights; second, the welfare of a given innovation is higher under competition that under monopoly; third, as just mentioned, innovations which occur only under an intellectual property regime are less valuable.

In addition to its 'stand-alone' uses, we also believe our model is valuable in its potential for integration into other innovation frameworks. In this paper, at least in relation to innovation, there is no downside to intellectual property rights and therefore, almost by assumption, an IP regime will outperform a no IP regime.[7] It would therefore be interesting to combine what we have here with more sophisticated models of the innovation process, for example one which incorporates cumulativeness. One of the main deficiencies of the cumulative innovation literature has been a lack of attention to the question of competition in the end product market – and how such competition changes with the IP regime.[8] Combining this paper's explicit modelling of imitation and competition in the end product market with a more sophisticated model of innovation would deliver a 'best-of-both-worlds' model, with an improved ability to capture both the benefits, and costs, of intellectual property rights.

4.1.1 Existing Literature

There are, of course, some papers in the existing literature which do allow for non-trivial imitation. For example Gallini (1992), allows patented innovations to be imitated for some fixed cost K. With free entry of imitators, K is then the maximum income achieved by an innovator who patents. Thus, in this model, imitation costs must be higher than innovation costs for innovation to occur.[9] In our model, by contrast,

with intellectual property rights. Similarly the welfare ratios is the level of welfare without intellectual property rights versus the level with.

[7]Rather what we are trying to investigate here is how wide the gap is. With perfect nonrivalry without intellectual property rights innovation is zero. We show that allowing for non-zero imitation, even if quite small, can dramatically change this result.

[8]For example, Bessen and Maskin (2006) assume in their model of cumulative innovation that, without intellectual property rights, each of the two firms receives some exogenously given share s of profits of that obtained with intellectual property rights. Meanwhile, Pollock (2006), following the approach of e.g. Denicolo (2000) and Bessen (2004) assumes that the IP regime only affects licensing and does not impact on the stand-alone value of the innovations.

[9]This is not precisely correct since Gallini allows for a firm not to patent – with non-patented inventions imitated at zero cost but only with some exogenous probability p_D. However, in this case (i.e. a firm does not patent) (a) there is no imitation cost – imitation either happens or it does not

imitation costs, both with and without intellectual property rights, may take any value (and without intellectual property rights are usually assumed to be less than innovation costs).

Other approaches include those based on locational models such as Waterson (1990) and Harter (1994) which both feature entry by a competing (imitative) firm within a horizontal product differentiation framework and focus on the impact of patent breadth on innovation and welfare.[10] This locational approach is obviously well-suited to considering imitation but is limited by the fact that it is extremely hard to endogenize entry. Both of the papers mentioned limit (imitative) entry to at most one firm. This makes it hard to analyze how changes in imitation cost impact on market structure and the innovator's rents. By contrast, we adopt a Stackelberg model of first-mover advantage. While this is obviously restrictive in other ways it allows us to tractably analyze equilibrium imitative entry.

Finally, Pepall and Richards (1994) also present a model which permits non-trivial imitation. Similar to our paper their model features Stackelberg competition with the innovator taking the role of the leader. However, their focus is on quality choice by the innovator and how imitation may lead to welfare losses due to inefficiently low choice of product quality. We, on the other hand, are more interested in exploring how variations in relative imitation cost impact on innovation, and how, incorporated into a model of the distribution of innovations at the aggregate level, this in turn can be used to examine the relative welfare performance of different regimes.

4.2 The Model

As should be clear from the survey of the empirical data above, in modelling imitation there are two basic directions in which to advance: imitation may be costly in terms of money or in terms of time.[11]

Here we shall confine ourselves to the case of imitation which is costly in terms of money and shall retain the assumption that it is costless in terms of time, i.e. instantaneous. Specifically, we adopt a model based on the Stackelberg model of quantity

with some exogenous probability; (b) IP rights are irrelevant.

[10]The impacts of patents is rather different in the two models. In Waterson (1990) it is an exclusion zone enforced via imperfect litigation (with fixed imitation costs) while in Harter (1994) the effect of a patent has a rather different dual effect: it makes imitation cheaper but the imitator must locate her product outside of the exclusion zone set by the patent.

[11]There are clearly other possibilities, for example imitation may be limited by the availability of skilled labour, or access to other necessary complementary assets (see e.g. Teece (1986)). However, these are both more complex to model and, we believe, of lesser importance than the main factors of time and money.

competition with multiple followers.[12] In our case, the first mover role is naturally taken by the developer of the original innovation whom we term the 'innovator', and the role of followers by 'imitators'. In the Stackelberg game the first mover advantage derives from the ability to commit to a particular output level before other players. Here, however, it is better to see the Stackelberg framework simply as a convenient method for modelling an advantage that derives from far more general sources, for example lead time, learning curve effects and the ability to put in place a marketing and sales operation (to take some of the items frequently cited in the empirical literature referred to in the introduction).

In all other respects firms are the same except for the fact that the innovator has different fixed costs from those of imitators. These fixed costs, both of the innovator and the imitators, are assumed to be non-zero – this along with the first-mover advantage is the key aspect of the model and again this assumption is based on the empirical evidence that was discussed above. There is no formal delay in innovation but the Stackelberg framework implicitly assumes the first-mover has time enough to commit to supply as much of the market as she wishes. Demand is taken to be linear with an inverse demand curve $p(q) = a - bq$. To summarize:

1. F_i the fixed cost of development for the innovator.

2. F_m the fixed cost of imitation which is assumed to be common across all imitators. Also define ϕ to be imitation cost as a proportion of innovation cost, so $\phi = F_m/F_i$. We assume that imitation cost is always less than innovation cost and that in the presence of intellectual property rights imitation does not occur (which could be interpreted as having infinite imitation cost).[13]

3. $c(q)$, marginal cost of production once the product is developed. It is assumed to be common between imitators and innovators (they both end up using the same technology), to be constant, and, without loss of generality, to be equal to zero.

4. Linear demand given by $p(q) = a - bq$

We have a slight variation on the classic two-stage model in which the sequence of actions can be considered as falling into three periods as follows:

[12]It could therefore be argued there is some temporal aspect in that the innovator is able to 'move' before imitators. However, there is no real imitation lag in the sense of a period of time in which the original innovator enjoys a monopoly of the relevant market.

[13]Note that this does not fit with the empirical data from Levin et al. (1987) where in several cases the costs of imitation exceeded those incurred by the innovator. Nevertheless, as the assumption greatly simplifies the analysis and incorporating the more complex reality would only strengthen our results, we feel warranted in proceeding as indicated.

1. An innovator decides whether to enter. If the innovator does enter then (s)he incurs a fixed cost, F_i, and develops a new product

2. Imitators decide whether to enter. If an imitator does enter (s)he incurs a fixed cost of F_m, and then has capacity to produce the new product.

3. Production occurs with price and quantities determined by Stackelberg competition in which the 'innovator' has the first-mover role and all imitators move simultaneously.

4.2.1 A Normalization

Define $k = \frac{a^2}{4b}$ so k is equal to half the area under the demand curve and therefore the level of monopoly profit. No agent's profits (innovator or imitator) can be greater than monopoly profits k. Hence let us simplify by normalizing all profits and fixed costs by dividing them by k – equivalent to setting k equal to 1 in the analysis below. Thus from now on when profits or fixed costs are discussed they should be taken not as absolute levels but as proportions of monopoly profits (itself equal to half of total potential welfare offered by the innovation). Formally:

$$\begin{aligned} f_i &= \tfrac{F_i}{k} \\ f_m &= \tfrac{F_m}{k} \end{aligned}$$

Note that, ϕ, the ratio imitation cost is also equal to the ratio of the normalized costs: $\phi = F_m/F_i = f_m/f_i$.

4.2.2 The Space of Innovations

In this model an innovation is specified by the tuple consisting of its 'innovation' cost and its 'imitation' cost: (f_i, f_m) (or equivalently (f_i, ϕ).[14] Innovation and imitation costs are non-negative, $f_i, f_m > 0$, and we have assumed that imitation costs are never more than innovation costs: $f_m \leq f_i$. Furthermore, it will never be optimal for an innovator to enter if $f_i > 1$, since the maximum possible profits from entering the market (k) are less than the cost of the innovation.

Thus, under the assumptions given and using normalized variables the space of innovations is $IS = \{(f_i, f_m) \in [0,1] \times [0,1] : f_m \leq f_i\} = \{(f_i, \phi) \in [0,1] \times [0,1]\}$.

[14]This conveniently allows us to visualize innovation space in a two dimensional graph (see the figures below for examples).

4.2.3 Policy Regimes and the Effect of Intellectual Property Rights

We will wish to consider different policy regimes. A given policy regime (R) has an associated model which will determines the costs and rents for the different agents and thereby defines some region in innovation space, IS, in which innovation occurs. It will also determine the welfare which each of those innovations generates.

In addition, a policy regime (R) will be taken to define a distribution of innovations over innovations over the innovation space IS which can be represented by some density function, say g^R. This function is primarily intended to capture information about the distribution of innovations at the aggregate level, for example industry or economy wide. This will be important because one cannot make decisions about the strength or presence of intellectual property rights on a firm-by-firm or technology-by-technology basis. Instead a policy-maker must set them at a very macro level – for example the length of patent protection is set by international treaty and must be the same across all patentable technologies. Even where there is choice, as in recent debates as to whether to extend patentability to software or copyright to perfumes, the decision must be made for an entire class of products displaying very substantial heterogeneity.[15]

In this paper we shall be interested in comparing and contrasting two particular regimes: that with intellectual property rights (e.g. patent or copyright) and that without. As just discussed, these regimes can differ both in their model (which determines whether a given innovation (f_i, f_m) occurs and the welfare it generates) and in the distribution of innovations over innovation space.[16] We focus on two distinct possibilities, with the first approach being the one we shall use by default:

1. Models differ, distributions the same. Specifically, under the no IP regime we use the Stackelberg model presented above. With IP we assume that all imitation is prohibited and that, as a result, the innovator makes monopoly profits.[17]

2. Models the same, distributions differ. Specifically, both regimes use the 'Stackelberg' model presented above but the distribution of innovations under no IP, g, is transformed to a new distribution g' under the IP regime. A graphical illustration of what this means is presented in Figure 4.1.

[15] A secondary purpose for the distribution function is to capture uncertainty by interpreting this function as representing the 'beliefs' of a policy-maker.

[16] In some ways allowing variation in the distribution of innovations is redundant since any variation in distribution could be incorporated as a difference in models. However, changes in distributions provide a simpler approach, that is less cumbersome in notation and more intuitive for understanding.

[17] This can be nested within our 'Stackelberg' model by restricting the number of imitators to be 0.

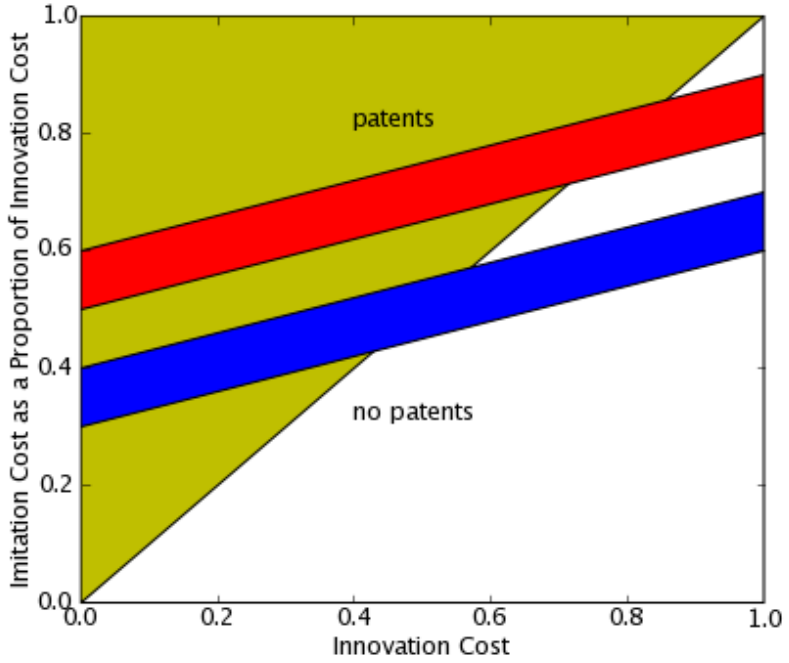

Figure 4.1: The Effect of Different Policy Regimes under the 'Breadth' approach. The light-shaded region above the diagonal indicates the area in which innovation occurs (the IP and no IP regime share a common model so this region is the same for both regimes). The lower band labelled 'no patents' indicated the distribution of innovations under the no IP regime while the upper band (labelled 'patents') indicates the new distribution of innovation with intellectual property rights (the implicit assumption here is that the introduction of intellectual property rights raises imitation costs by some fixed proportion leaving innovation costs unchanged).

For future reference we shall label the first case the 'Zero Imitation' (ZI) approach to modelling intellectual property rights (and label the associated regime the 'Zero Imitation' regime) and the second the 'Breadth' (BR) approach to modelling intellectual property rights.

4.3 Solving the Model

We solve by recursing backwards through the game. First, in Proposition 4.3.1, we determine the solution to the Stackelberg model of price competition in the final product market assuming a fixed, given number of imitators. Next we solve for the equilibrium number of imitators using the zero-profit condition generated by the assumption of free entry. This gives the number of imitators as a function of the imitation cost f_m. Using this, we can determine the innovator's expected gross profits as a function of the number of imitators (and hence imitation cost f_m). If these profits exceed the innovation cost, f_i then the innovator would enter and the innovation occurs – otherwise it does not. We summarize the results in Propositions 4.3.3 and 4.3.4, which details the set of innovations occurring in equilibrium.

Proposition 4.3.1. *Let n be the exogenously given number of imitators. The solution to the Stackelberg model of competition by quantify is as follows where k is defined as above to equal $a^2/4b$ ('i' subscripts are on variables related to the innovator and 'm' subscripts are on variables related to an imitator):*

$$\begin{aligned}
q_i &= \frac{a}{2b} \\
q_m &= \frac{a}{2b(n+1)} \\
\textit{Total output} &= Q = \frac{a(2n+1)}{2b(n+1)} \\
p &= a - bQ = \frac{a}{2(n+1)} \\
\textit{Gross profits of an innovator} &= \Pi_i = \frac{k}{n+1} \\
\textit{Gross profits for an imitator} &= \Pi_m = \frac{k}{(n+1)^2} = \frac{\Pi_i}{n+1}
\end{aligned}$$

Proof. Omitted (the solution to the Stackelberg model is well-known). □

Proposition 4.3.2. *Imposing a zero net profit condition on the basis of free entry as an imitator, the number of imitators, n^e is as follows:*

- *Non-integer n:* $n^e = \sqrt{\frac{1}{f_m}} - 1$
- *Integer n:* $n^e = \max\{n \in \mathbb{Z} : f_m \leq \frac{1}{(n+1)^2}\}$

Proof. Allowing non-integer n we solve $\Pi_m = F_m$. This gives:

$$n^e = \sqrt{\frac{k}{F_m}} - 1 = \sqrt{\frac{1}{f_m}} - 1$$

Restricting to integer n we require the n such that $\Pi_m \geq F_m$ but with $n+1$ imitators $\Pi_m < F_m$. Substituting for Π_m gives the condition. □

Proposition 4.3.3. *Allowing the number of imitators to take non-integer values then an innovation (f_i, f_m) occurs if $f_m \geq f_i^2$ ($\phi \geq f_i$). Thus, the set of innovations which occur is given by:*

$$A^c = \{(f_i, f_m) \in IS : f_m \geq f_i^2\} = \{(f_i, \phi) \in IS : \phi \geq f_i\}$$

Proof. Innovation only occurs if expected (net) profits are positive, that is $\Pi_i \geq f_i$. Substituting for the LHS using our value for the number of imitators from Proposition 4.3.2 gives the condition:

$$f_m \geq f_i^2$$

□

Proposition 4.3.4. *Restricting the number of imitators to integer values the set of innovations that occur is:*

$$A^{int} = \cup_{n=0}^{\infty}\{(f_i, f_m) \in IS : \frac{1}{n^2} \geq f_m > \frac{1}{(n+1)^2}, f_i \leq \frac{1}{n+1}\}$$

Proof. Direct from Proposition 4.3.2 □

Remark 4.3.5. Note the substantial difference between the two situations (non-integer and integer numbers of imitators). For example, with integer-only number of imitators, $f_m \geq \frac{1}{4} \Rightarrow n = 0$ and all innovations with $f_i \leq 1$ are realized, a very different outcome to that with continuous number of imitators. We return to this theme below, in Proposition 4.3.7.

In this model an innovation is defined by a pair (f_i, f_m) giving its innovation and imitation cost. We can therefore visualise potential innovations in a two dimensional graph of innovation/imitation cost space. In particular, we can summarize the results of the previous propositions in Figure 4.2. In this diagram the light-shaded (yellow) region is that in which innovations occur with non-integer numbers of imitators permitted,

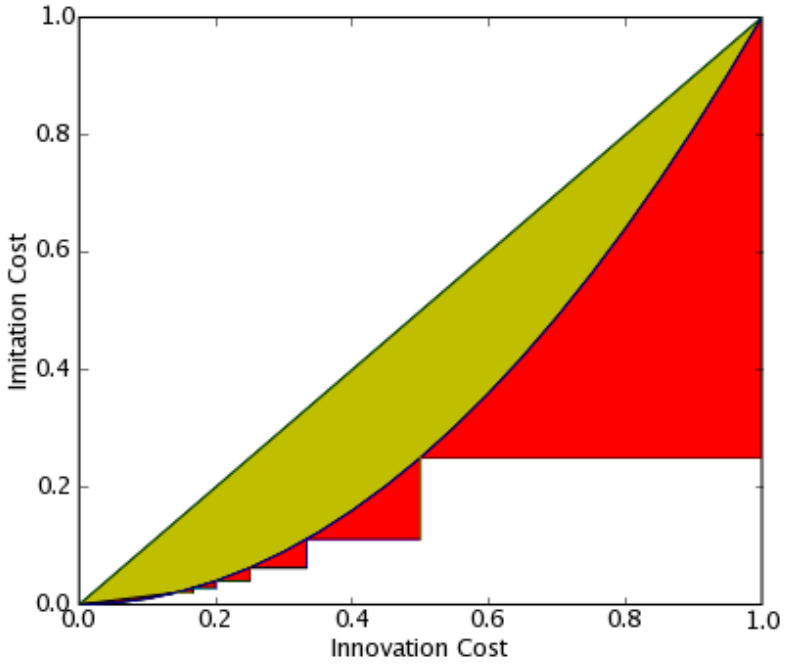

Figure 4.2: Innovations which occur without intellectual property rights (light shaded region: non-integer numbers of imitators allowed; dark-shaded extra innovations occurring when restricted to integer numbers of imitators).

while the innovations in the dark-shaded (red) and light-shaded region occur when restricting to integer numbers of imitators. (The region above the diagonal should be ignored since we are assuming that imitation cost is always less than innovation cost).

While the preceding diagram is entirely correct as it stands, it will be useful to visualize the same data in a slightly different manner. We do this by replacing imitation cost by 'proportional' imitation cost (ϕ) – i.e. imitation cost as a proportion of innovation cost. Under our assumption that imitation cost is always less that innovation cost this means that we now have a constant range, [0,1], for 'proportional' imitation cost at all levels of innovation cost and, in visual terms, we have a uniform level of innovation per unit of innovation cost. This is shown in Figure 4.3 which is simply a re-rendering of Figure 4.2 using proportional innovation cost.

Proposition 4.3.6. *With intellectual property rights (zero imitation) all innovations in IS occur and $A^{IP} = IS$*

Proof. We have assumed that with intellectual property rights no imitation is possible hence an innovation occurs if and only if innovation costs are less than 1. □

Thus, with IP, all of the area under the 45 degree line in Figure 4.2 and all of the area in Figure 4.3 would be shaded.

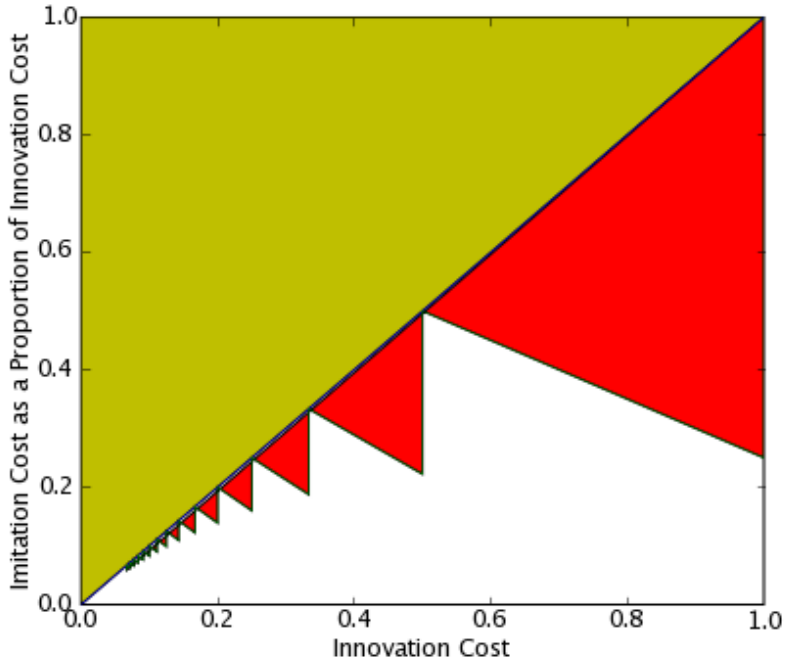

Figure 4.3: Innovations which occur without intellectual property rights (light shaded region: non-integer numbers of imitators allowed; dark-shaded extra innovations occurring when restricted to integer numbers of imitators).

Returning to our theme of the difference between allowing continuous and integer numbers of imitators, we have:

Proposition 4.3.7. *Assuming a uniform distribution over the space of innovations shown in Figure 2 (this corresponds to calculating area), that is with density function* $g(f_i, \phi) = 1$*, the ratio of innovation without intellectual property rights to that with intellectual property rights is: 50% (non-integer n), 72% (integer n).*

Proof. See appendix. □

Thus restricting to integer n increases the amount of innovation by nearly 50% and much of this extra innovation occurs at the higher levels of innovation and imitation cost when the number of imitators in the integer case will be low (zero, one or two). Despite, this difference in the remainder of the paper we shall, by default, focus on the case of continuous n. This is for two reasons. First, especially when performing integrations to obtain welfare totals, the continuous case is much easier to use. Second, as just shown, restricting to integer n will only strengthen our results regarding the relative performance of a no IP regime. Thus, any result we obtain for continuous numbers of imitators, will hold a fortiori for discrete number of imitators.

4.4 Welfare and Policy

From a policy perspective what really matters is the utility generated by innovation not how much innovation occurs. If the welfare from innovations realized without intellectual property rights differ systematically from those that are not or the welfare generated by a given innovation differs under the two regimes then welfare outcomes will differ from innovation levels.

Let R and S denote two distinct policy regimes. Define:[18]

$$\begin{aligned} W^R(f_i, \phi) &= \text{Welfare under regime R from innovation}(f_i, \phi) \\ \Delta W_S^R(f_i, \phi) &= W^R(f_i, \phi) - W^S(f_i, \phi) \end{aligned}$$

4.4.1 Welfare Per Innovation

Take R to be the no IP regime (NIP) and S to be the IP/zero imitation regime (ZI). Recall that under ZI all innovation in the innovation space, IS, occur. Let A denote the region in which innovation occurs under NIP, then we have:

Proposition 4.4.1. *The difference in welfare generated by an innovation (f_i, ϕ) under the no IP regime (NIP) compared to the zero imitation regime (ZI) is:*

$$\Delta W^{NIP}(f_i, f_m) = \begin{cases} \frac{n^{e2}}{2(n^e+1)^2}, & (f_i, \phi) \in A \\ -W^{ZI}(f_i, \phi), & (f_i, \phi) \in IS - A \end{cases}$$

In particular, when the innovation is in A – and therefore occurs under both regimes – this difference is always non-negative and the no IP regime generates more welfare than the zero imitation regime.

Proof. See appendix. □

The ΔW term captures the fact that, for a *given innovation*, the welfare generated by it differs between the two regimes. This difference is driven by two distinct, and contrary, effects. First, no intellectual property rights leads to greater competition. This transfers rents from producers to consumers and reduces the deadweight loss because total output expands. Second, with imitation there is greater entry which means total fixed costs expended for a given innovation are higher due to the greater number of producers. In this model, the first effect outweighs the second (conditional, of course, on the innovation still being produced without intellectual property rights).[19]

[18] Note that if the innovation (f_i, ϕ) does not occur under regime R then $W^R(f_i, \phi) = 0$.

[19] This result has a simple, intuitive, basis. Under a Stackelberg model of quantity competition the

4.4.2 A Single Technology With Observable Costs

Corollary 4.4.2. *Assume costs are precisely observable by a regulator. If IP is represented by the 'Zero Imitation' regime, the optimal policy rule is to grant intellectual property rights if and only if the square of innovation costs (as a proportion of monopoly profits) is larger than imitation costs (also as a proportion of monopoly profits), that is:* $f_i^2 > f_m$.

Proof. Our previous result shows that welfare without intellectual property rights is greater than than with intellectual property rights (Zero Imitation) if and only the innovation occurs without intellectual property rights. Thus the 'square' rule follows directly from from our result on innovation as described in Propositions 4.3.2 and 4.3.3. □

This 'square' rule is we believe a novel result in the literature. While its convenient form is clearly specific to the Stackelberg-type model we have adopted, as we show below, the point that the 'allowable' imitation cost falls (that is the minimal imitation cost such that innovation still occurs) as innovation cost falls is a general one.

We also note that if IP is represented by the 'Breadth' regime rather than a 'Zero Imitation' regime a very similar result still obtains. To be precise, assuming that an increase in breadth acts to increase imitation costs leaving innovation costs unchanged, then, given an innovation with costs (f_i, f_m) (under no IP), the optimal policy rule consists in setting the breadth of the IP right such that if f'_m is the new imitation cost (under IP) then $f'_m = f_i^2$.

4.4.3 A Distribution of Innovations

The results of the previous section are certainly valuable, however, they suffer from two significant drawbacks if intended for use by regulators in the real-world. First a regulator usually lacks precise information about innovation and imitation costs (at least ex-ante). Second, and more importantly, as discussed above in Section 4.2.3, a policy-maker cannot make decisions about the strength or presence of intellectual property rights on a technology-by-technology basis. Instead decisions about the existence, and strength, of such rights must be taken at a much more aggregate level.[20]

output of the leader (the innovator) stays fixed at the monopoly level. Thus, the income used to cover imitators' fixed costs must always come from output expansion. Hence, though imitative entry does increase fixed costs those fixed costs are always less than the increase in surplus arising from the output expansion.

[20]And this is not simply for informational reasons but because of the need to be compatible with existing norms and agreements. For example, an international treaty (TRIPS) sets down a minimum length for patent protection and mandates that it must be the same across all patentable technologies.

Thus, in this section we extend our welfare analysis to the aggregate, industry or economy-wide, level by incorporating the distribution of innovations. Using the notation set out in the Section 4.2.3 above we encapsulate the distribution of innovations under a given regime, R, in a probability distribution function g^R defined over the space of innovations IS. Extending our existing notation we have:

$$\begin{aligned} W^R(X) &= \text{Welfare from region X under regime R} = \int_X W^R(f_i, f_m) g^R \\ \Delta W_S^R(X) &= W^R(X) - W^S(X) \end{aligned}$$

We shall focus again on the no IP (NIP) and zero imitation (ZI) IP regime. As stated in Section 4.2.3, we assume these share the same distribution of innovations. We shall therefore drop the superscript and simply use g for this distribution. Recall also that, under the zero imitation regime, all innovation in IS takes place. Let A be the region in which innovation takes place under no IP and define $B = IS - A$, that is, the set of innovations not in A. Then:

$$W^{ZI} = W^A(ZI) + W^{ZI}(B) \tag{4.1}$$

$$W^R = W^R(A) + W^R(B) = W^{ZI}(A) + \Delta W_{ZI}^R(A) \tag{4.2}$$

The second equation illustrates how we may break up the welfare under regime R. First, note that the welfare from region B, $W^R(B)$ is zero since, by definition, no innovation occurs in that region. Turning to region A, we may divide welfare that we would get in the case of zero imitation (the first term) plus the difference between that level and the level of welfare in regime R: ΔW.

This allows us to distinguish between three effects that operate with respect to differences in welfare. First, less innovation occurs under no IP compared to Zero Imitation. Second, is the fact, already mentioned, that, assuming an innovation occurs under both regimes, it will generate more welfare under no IP than under Zero Imitation. This is captured in the ΔW term. Third, is the fact that innovation fixed costs may differ systematically between regions A and B (A is the region in which innovation occurs under both regimes while B is everything else). This will materialize in the relative sizes of $W(A)$ and $W(B)$. We illustrate these effects with a simple example where innovations are uniformly distributed:

Proposition 4.4.3. *Assuming a uniform distribution over the space of innovations as shown in Figure 2, that is with density function $g(f_i, \phi) = 1$, welfare levels are*

as follows (where NIP indicates a regime without intellectual property rights and the number of imitators may take non-integer values):

$$W^{ZI}(A) = \tfrac{7}{12}, \textit{ average welfare density} = \tfrac{7}{6} \tag{4.3}$$

$$W^{ZI}(B) = \tfrac{5}{12}, \textit{ average welfare density} = \tfrac{5}{6} \tag{4.4}$$

$$\Delta W(A)^{NIP} \approx \tfrac{2}{12}, \textit{ average welfare density} = \tfrac{2}{6} \tag{4.5}$$

Proof. See appendix. □

Thus, the ratio of welfare without intellectual property rights to a situation in which they are present is **75%**. Comparing this with the results of Proposition 4.3.7 we see that a regime without intellectual property rights while only having half the level of innovation delivers three quarters of the welfare achieved with intellectual property rights. Furthermore, we see that the third effect mentioned above, that is the systematic difference in the fixed cost of innovation, is a significant driver of these results. For example, if we were to assume that ΔW were zero, that is the welfare generated by innovations under no IP and IP were the same, we would still have a welfare ratio of 58% – the same gain if there under the converse assumption of no difference in fixed costs but only differences in per innovation welfare yields.

To give another illustration, consider now the question of uncertainty. Suppose a policy-maker knows precisely the proportional imitation costs but has complete uncertainty regarding innovation costs (so the policy-makers belief are represented by a uniform distribution over the possible values).[21]

Proposition 4.4.4. *Assuming a uniform distribution of innovation costs if imitation costs are more than 70% of innovation costs then welfare is higher without intellectual property rights.*

Proof. See appendix. □

Turning to the case where innovation costs are known with certainty but there is complete uncertainty regarding imitation costs one has a similar result:

Proposition 4.4.5. *Assuming a uniform distribution of proportional imitation costs, if innovation costs are less than 20% of total potential monopoly profits then welfare is higher without intellectual property rights.*

Proof. See appendix. □

[21] For example, the data provided in Levin et al. (1987) provide information on proportional imitation costs but nothing on the cost of innovation itself.

4.5 The General Case

The quantitative results obtained above must clearly be specific to assumptions regarding the underlying model and distribution of innovations. However, the basic point that welfare proportions will always be systematically higher than innovation proportions (even if we ignore deadweight loss) holds in general.

Recall that an innovation is specified by the tuple (f_i, f_m) (or equivalently (f_i, ϕ) and that (using normalized variables) the space of innovations is then $IS = \{(f_i, f_m) \in [0,1] \times [0,1] : f_m \leq f_i\} = \{(f_i, \phi) \in [0,1] \times [0,1]\}$.

Now, any given regime R (with associated model of innovation and imitation M^R) will define some region in IS in which innovation occurs. Following previous convention we will denote this region by A. We make the mild assumptions that:

Assumption 4.5.1. Suppose the innovation $I^1 = (f_i^1, f_m^1) \in A$ then:

1. Any other innovation with the same imitation cost but lower innovation cost occurs under R. Formally: $\forall f_i \leq f_i^1, (f_i, f_m^1) \in A$.

2. Any other innovation with the same innovation cost but higher imitation cost occurs under R. Formally: $\forall f_m \geq f_m^1, (f_i^1, f_m) \in A$.

How can we characterise this region, A, in which innovation occurs under regime R? Define $h(f_i)$ as the infinum of all innovations with innovation cost f_i that are in A:

$$h(f_i) = inf\{f_m : (f_i, f_m) \in A\}$$

Let us assume (without loss of generality) that $h(f_i) \in A$.

Proposition 4.5.2. *The area in which innovation occurs A is given as follows:*

$$A = \{(f_i, f_m) \in IS : f_m \geq h(f_i)\}$$

Furthermore, h is a non-decreasing function.

Proof. The first part follows directly from Assumption 4.5.1.2 combined with the definition of the supremum h. To show that h is non-decreasing suppose not, that is that there exists $f_i^1 < f_i^2$ such that $f_m^1 = h(f_i^1) > h(f_i^2) = f_m^2$. By Assumption 4.5.1.1 $(f_i, f_m^2) \in A, \forall f_i < f_i^2$ which implies, in particular, $(f_i^1, f_m^2) \in A$, but $f_m^2 < f_m^1$ which implies $h(f_i^1) \leq f_m^2 < f_m^1 = h(f_i^1)$ which is a contradiction. □

Definition 4.5.3. Given a regime R recall that I^R is the amount of innovation occurring under R and W^R the total amount of welfare. Then given two different regimes, R, S, define:

1. IR(R,S) = Innovation Ratio of R to S = the ratio of innovation under R to innovation under S

2. WR(R,S) = Welfare Ratio of R to S = the ratio of welfare under R to welfare under S

Proposition 4.5.4 (Welfare Ratio is higher than Innovation Ratio)**.** *Take a general regime R and a corresponding zero imitation (ZI) regime (so the ZI regime shares the same distribution of innovations as R). Assume that welfare from a given innovation (if it occurs under both regimes) generates at least as much welfare under R as under ZI:*

$$W^R(f_i, f_m) \geq W^{ZI}(f_i, f_m)$$

Then the welfare ratio *of R compared to zero imitation ZI will be greater than or equal to the* innovation ratio *of R compared to zero imitation (ZI). Furthermore, the inequality is strict if there is any innovation which occurs under R and there are some innovations which occur under ZI but not under R. That is:*

$$WR(R, ZI) \geq IR(R, ZI)$$

Proof. See appendix. □

Remark 4.5.5. Note that this result holds even if there are *no* deadweight losses, that is the welfare generated under R per innovation is the same as under ZI. Hence, this proposition establishes in great generality the point made earlier that the narrowing of the differential between the no IP and IP regime when moving from innovation to welfare was driven not simply by the well-known welfare-benefits of greater competition but also by systematic differences in the average of costs of innovations occurring with and without IP.

4.6 Conclusion

In this paper we have presented a simple model of innovation with imitation. We have shown that when imitation is costly and there is some form of first mover advantage the initial innovator may still be able to garner sufficient rents to cover the fixed cost of development even though not enjoying a pure monopoly. As discussed in the introduction, there is a great deal of empirical support for believing imitation costs and first mover advantage are important. This paper demonstrates that these concerns can be analyzed simply and tractably, and, that doing so, generates important new insights

– most significantly that ignoring them may overstate the importance of intellectual property rights.

Here innovations are specified by a tuple consisting of the 'innovation' cost and the 'imitation' cost (the innovation cost being the cost to the first developer of the product). Using our Stackelberg-based model of first-mover advantage we obtained a precise description of which innovations would occur with imitation (that is, without IP rights). The formula took a particularly simple form which we dubbed the 'square' rule because it stated that innovations occurred if and only if (normalized) imitation cost was greater than the square of (normalized) innovation cost (we normalized by dividing by the potential monopoly profit so that all costs were in the range $[0, 1]$). While this particular formula must necessarily be dependent on the precise structure of the underlying model, the basic point that 'allowable' imitation costs fall with innovation cost is, we believe, a very general one – one, furthermore, which has received scant notice in previous literature.

Next we turned to a consideration of welfare and its implication for policy. We first showed that the 'square' rule carried over from innovation to welfare. This has important policy consequences. For example, if the ratio of imitation costs to innovation costs are the same in two industries but the (normalized) cost of innovation differs, then the impact of intellectual property rights in the two industries will be very different. Specifically, in the industry with lower innovation costs, the benefits of IP will be much lower (and could even be negative). This result illustrates how the impact of IP may vary in a systematic way across industries. In particular there will be industries in which intellectual property rights are necessary – and industries where they are not, and this paper presents one basis for a taxonomy to determine which is which.

However, it is rare that a policy-maker knows precisely the innovation and imitation costs for a given technology. Furthermore, it is, in practice, impossible for a policy-maker to set the level of IP on a technology, or even industry-by-industry basis. Hence, the next step was to extend our analysis to consider the case where there is a *distribution* of innovations – this distribution can be taken to represent either beliefs, or a collection of potential innovations at the industry or economy-wide level.

Comparing regimes without and with intellectual property rights we showed that the welfare ratio is systematically higher than the innovation ratio. Moreover, it was demonstrated that this is not simply for the familiar reason that, conditional on the innovation being made, greater competition without intellectual property rights leads to in increased output and lower deadweight losses. Rather, there was the additional factor, namely that the set of innovations occurring under an IP regime are, on average, less socially valuable because they have higher fixed costs of creation.

Finally, we note that there are a variety of way in which the present work could be extended. One could, for example, introduce a 'race' for the innovation in standard manner. This would allow for multiple firms at the innovation stage competing to produce the original innovation. This could be extended so that failed innovators can be imitators at the second stage.

On a separate point, one distinctive feature of this model is that intellectual property rights always lead to maximal innovation. In a more complex model, for example one involving cumulative innovation, this might no longer be the case. There are a variety of approaches that could be taken to integrate such dynamics and investigating these options would be one of most important improvements to the model that could be made.

Another option, which has already been mentioned briefly, is to have a richer model of imitation delay. Similarly, allowing for types of imperfect competition other than Stackelberg would also be a valuable extension. For example, the models of Waterson (1990) and Klemperer (1990) both provide for product differentiation and these models could be adapted to provide a richer and more realistic model of imitation in the presence – and absence – of intellectual property rights.

4.A Proofs of Propositions

Proof of Proposition 4.3.7. A uniform distribution of innovation corresponds to the standard euclidean measure over IS, which in turns corresponds to calculating areas in Figure 2. With intellectual property rights no imitation is permitted so all the innovations in the figure occur (total area of the figure is 1). Thus to calculate the proportions of innovation occurring without intellectual property rights we need to calculate the size of the dark-shaded and light-shaded areas as proportion of the entire figure.

For continuous n we consider the light-shaded region. This, clearly, has area equal to 1/2.

Restricting to integer n we need to add to this the area of the dark-shaded (red) region. The area of the dark-shaded (red) region is made up of a series of similar triangles. The nth triangle (working down from the largest) has area:

$$0.5 \cdot b \cdot h = 0.5 \cdot (\frac{1}{n} - \frac{1}{n+1}) \cdot (\frac{1}{n} - \frac{n}{(n+1)^2})$$

Thus total area of dark-shaded (red) region is:

$$0.5\sum_{1}^{\infty}(\frac{1}{n}-\frac{1}{n+1})\cdot(\frac{1}{n}-\frac{n}{(n+1)^2}) = 0.5\cdot(\sum\frac{1}{n^2}-\sum\frac{1}{n(n+1)}-\sum\frac{1}{(n+1)^2}+\sum\frac{n}{(n+1)^3})$$

All of these sums are simple except for the third. For this one approximate as follows:

$$\sum\frac{n}{(n+1)^3} \approx \sum_{1}^{99}\frac{n}{n+1}^{3} + \int_{99}^{infty}\frac{1}{(x+1)^2} = 0.432976 + 0.01 = 0.4430$$

Substituting this gives the dark-shaded (red) region's total area as:

$$0.5\cdot(\sum\frac{1}{n^2}-\sum\frac{1}{n(n+1)}-\sum\frac{1}{(n+1)^2}+\sum\frac{n}{(n+1)^3}) = 0.5\cdot((1+X)-1-X+0.4430) = 0.2215$$

Thus total area of light-shaded and dark-shaded region is $0.5 + 0.2215 \approx 0.72$.

□

Proof of Proposition 4.4.1. First let us determine the welfare arising from a given innovation. If there are n imitators we have that consumer surplus (CS) and producer surplus (PS) are as follows:

$$CS(f_i, f_m) = 0.5 \cdot (a-p) \cdot q = \tfrac{(2n+1)^2}{2(n+1)^2} \tag{4.6}$$

$$PS(f_i, f_m) = \Pi_i - f_i + n \cdot (\Pi_m - f_m) = \tfrac{1}{n+1} - f_i \tag{4.7}$$

Note that we have used the fact that, with continuous n, the zero profit condition implies $\Pi_m = f_m$. Summing to get total welfare we have:

$$W(f_i, f_m) = CS + PS = \frac{(2n+1)^2}{2(n+1)^2} + \frac{1}{n+1} - f_i$$

Now in a ZI regime $n = 0$ so:

$$W^{ZI} = \frac{3}{2} - f_i$$

Thus,

$$\Delta W(f_i, f_m) = W^R(f_i, f_m) - W^{ZI}(f_i, f_m) \tag{4.8}$$
$$= (\tfrac{(2n+1)^2}{(n+1)^2} + \tfrac{1}{n+1} - f_i) - (\tfrac{3}{2} - f_i) \tag{4.9}$$
$$= \tfrac{n^2}{2(n+1)^2} \tag{4.10}$$

□

Proof of Proposition 4.4.3. To calculate total welfare for region X we integrate welfare per innovation, $W(f_i, f_m)$, over X.

$$W^{ZI}(A) = \frac{1}{2}(\frac{3}{2} - \text{avg over A}(f_i)) = \frac{3}{4} - \frac{1}{2}\frac{1}{3} = \frac{7}{12}$$

$$W^{ZI}(B) = \frac{1}{2}(\frac{3}{2} - \text{avg over B}(f_i)) = \frac{3}{4} - \frac{1}{2}\frac{2}{3} = \frac{5}{12}$$

Calculating ΔW is slightly more complicated:

$$\Delta W(A) = \int_A \frac{n^2}{2(n+1)^2} = \int_0^1 \int_{f_i}^1 d\phi df_i$$

Recall that:

$$\phi = \frac{f_m}{f_i} \tag{4.11}$$
$$n+1 = \frac{1}{\sqrt{f_m}} \Rightarrow \frac{n^2}{(n+1)^2} = 1 - 2\sqrt{f_m} + f_m \tag{4.12}$$

Thus, substituting f_m for ϕ as well as for n we have:

$$\Delta W(A) = 0.5 \int_0^1 \frac{1}{f_i} \int_{f_i^2}^{f_i} 1 - 2\sqrt{f_m} + f_m df_m df_i$$

Working through the first integration gives:

$$\Delta W(A) = 0.5 \int_0^1 1 - \frac{4\sqrt{f_i}}{3} - \frac{f_i}{2} + \frac{4f_i^2}{3} - \frac{f_i^3}{2} df_i = \frac{13}{72} \approx \frac{1}{6}$$

□

Proof of Proposition 4.4.4. We need to determine welfare at a particular level of ϕ (imitation cost as a proportion of innovation costs) assuming a uniform distribution of innovation costs under an IP (zero imitation) and no IP regime. Proceeding as above but making all welfare calculations a function of ϕ we have:

$$W^{ZI}(A)(\phi) = \tfrac{1}{2}(3\phi - \phi^2) \tag{4.13}$$

$$W^{ZI}(B)(\phi) = 1 - \tfrac{1}{2}(3\phi - \phi^2) \tag{4.14}$$

$$\Delta W^{NIP}_{ZI}(A)(\phi) = \tfrac{1}{2}(\phi - \tfrac{4}{3}\phi^2 + \tfrac{\phi^3}{2}) \tag{4.15}$$

The difference in welfare between a regime without IP compared to one with is $\Delta W(\phi) = W^{NIP}(\phi) - W^{ZI}(\phi)$. Thus to determine the cut-off point, α say, such that for all $\phi \leq \alpha$ the no IP regime is preferable we simply need to solve:

$$\Delta W(\phi) = 0$$

(Note that ΔW is an increasing function of ϕ so the solution will be unique and that $\Delta W(0) < 0$ and $\Delta W(1) > 0$ so a solution will exist).

Proceeding numerically we obtain a figure of $\alpha = 0.704 \approx 0.7$. □

Proof of Proposition 4.4.5. We proceed as in the previous proof though this time focusing on welfare at a particular level of f_i (innovation cost as a proportion of potential monopoly profit) assuming a uniform distribution of proportional imitation cost under an IP (zero imitation) and no IP regime. Making all welfare calculations a function of f_i we have:

$$W^{ZI}(A)(f_i) = (\tfrac{3}{2} - f_i)(1 - f_i) \tag{4.16}$$

$$W^{ZI}(B)(f_i) = (\tfrac{3}{2} - f_i)f_i \tag{4.17}$$

$$\Delta W^{NIP}_{ZI}(A)(f_i) = \tfrac{1}{2}(1 - \tfrac{4\sqrt{f_i}}{3} - \tfrac{f_i}{2} + \tfrac{4f_i^2}{3} - \tfrac{f_i^3}{2} \tag{4.18}$$

The difference in welfare between a regime without IP compared to one with is $\Delta W(f_i) = W^{NIP}(f_i) - W^{ZI}(f_i)$. Thus to determine the cut-off point, α say, such that for all $f_i \leq \alpha$ the no IP regime is preferable we simply need to solve:

$$\Delta W(f_i) = 0$$

(Note that ΔW is a decreasing function of f_i so the solution will be unique and that $\Delta W(0) > 0$ and $\Delta W(1) < 0$ so a solution will exist).

Proceeding numerically we obtain a figure of $\alpha = 0.191 \approx 0.2$. □

Proof of Proposition 4.5.4. Claim: Assume the innovation $(f_i^1, f_m^1) \in A$. Then for any regime X if $f_i < f_i^1, W^X(f_i^1, f_m^1) > W^X(f_i, f_m^1)$.

Proof of Claim. Innovation cost is a sunk cost and the original innovation (f_i^1, f_m^1) is in A (and so occurs under either regime). Then reducing the cost of innovation has no effect on the behaviour of the innovator and as imitation cost are unchanged the solution of the model in terms of price, output etc must be the same. As a result Consumer Surplus must be unchanged and the only change to producer surplus comes from a reduction in the innovator's cost (which increases producer surplus). The claim follows. □

Under ZI all innovations in IS occur. Let A be the region of IS in which innovations occur under R. Let g be the probability distribution function on IS describing the distribution of innovations over the space. Define H as the inverse to h: $H = h^{-1}$. Pick a given proportional imitation cost f_m then it is sufficient to prove the result focusing on a single slice of innovation space at f_m. That is, if we can show that just looking at innovations with imitation cost f_m that the welfare ratio is higher than the innovation ratio then the result must hold when looking at the whole space of innovations.

Define $I^X(f_m), W^X(f_m)$ to be the innovation and welfare levels under the regime $X = R, ZI$ when restricting to innovations with imitation cost f_m. So considering the innovation ratio we have:

$$\begin{aligned}
\text{Innovation Ratio at } f_m &= \frac{I^R(f_m)}{I^{ZI}(f_m)} \\
I^R(f_m) &= \int_0^{H(f_m)} g(f_i, f_m) df_i \\
I^{ZI}(f_m) &= \int_0^1 g(f_i, f_m) df_i
\end{aligned}$$

Turning to welfare, by the Claim above for $f_i^1 \leq H(f_m) \leq f_i^2$ we have $W^{ZI}(f_i^1, f_m) \geq W^{ZI}(H(f_m), f_m) \geq W^{ZI}(f_i^2, f_m)$. Then for some C_1, C_2 with $C_1 > 1 > C_2$ we have:

$$\begin{aligned}
W^{ZI}(f_m) &= \int_0^1 W^{ZI}(f_i, f_m) g df_i \\
&= \int_0^{H(f_m)} W^{ZI}(f_i, f_m) g df_i + \int_{H(f_m)}^1 W^{ZI}(f_i, f_m) g df_i \\
&= C_1 W(H(f_m), f_m) \int_0^{H(f_m)} g df_i + C_2 W(H(f_m), f_m) \int_{H(f_m)}^1 g df_i \\
&\leq C_1' \left(\int_0^{H(f_m)} g df_i + \int_{H(f_m)}^1 g df_i \right) \\
&= C_1' I^{ZI}(f_m)
\end{aligned}$$

Note that the inequality is strict if there are innovations both in A and outside of A, that is $\exists f_i^1 < H(f_m) < f_i^2$ with $g(f_i^j, f_m) > 0, j = 1, 2$.

Now by assumption for any $(f_i, f_m) \in A$ (i.e. with $f_i \leq H(f_m)$), $W^R(f_i, f_m) \geq W^{ZI}(f_i, f_m)$. Thus,

$$\begin{aligned} W^R(f_m) &= \int_0^{H(f_m)} W^R(f_i, f_m) g df_i \\ &\geq \int_0^{H(f_m)} W^{ZI}(f_i, f_m) g df_i \\ &= C_1' \int_0^{H(f_m)} g df_i \\ &= C_1' I^R(f_m) \end{aligned}$$

Hence we have that the Welfare ratio of R to ZI at f_m (with the inequality being strict under the condition previously stated):

$$\begin{aligned} \text{Welfare Ratio}(f_m) &= W^R(f_m)/W^{ZI}(f_m) \\ &\geq C_1' I^R(f_m)/C_1' I^{ZI}(f_m) \\ &= I^R(f_m)/I^{ZI}(f_m) \\ &= \text{Innovation Ratio}(f_m) \end{aligned}$$

□

Bibliography

A. Arundel. Patents in the Knowledge-Based Economy. *Beleidstudies Technology Economie*, 37:67–88, 2001.

James Bessen. Hold-up and Patent Licensing of Cumulative Innovations with Private Information. *Economics Letters*, 82(3):321–326, 2004.

James Bessen and Eric Maskin. Sequential Innovation, Patents, and Innovation. NajEcon Working Paper Reviews 321307000000000021, www.najecon.org, May 2006.

W. Cohen, R. Nelson, and P. Walsh. Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not), 2000.

Victor Denicolo. Two-Stage Patent Races and Patent Policy. *Rand Journal of Economics*, 31:488–501, 2000.

Giovanni Dosi. Sources, Procedures, and Microeconomic Effects of Innovation. *Journal of Economic Literature*, pages 1120–1171, 1988.

Nancy Gallini. Patent Policy and Costly Imitation. *Rand Journal of Economics*, 23 (1):52–63, 1992.

Bronwyn Hall. Business Method Patents, Innovation and Policy, 5 2003.

John F. R Harter. The Propensity to Patent with Differentiated Products. *Southern Economic Journal*, 61(1):195–201, July 1994. ISSN 00384038.

Hugo A Hopenhayn and Matthew F Mitchell. Innovation Variety and Patent Breadth. *The RAND Journal of Economics*, 32(1):152–166, 2001. ISSN 07416261.

Paul Klemperer. How Broad Should the Scope of Patent Protection Be? *RAND Journal of Economics*, 21(1):113–130, 1990.

Richard Levin, A. Klevorick, R. Nelson, S. Winter, R. Gilbert, and Z. Griliches. Appropriating the Returns from Industrial Research and Development. *Brookings Papers on Economic Activity*, 3:783–831, 1987.

Edwin Mansfield. How Rapidly Does New Industrial Technology Leak Out? *Journal of Industrial Economics*, 34(2):217–223, 1985.

Peter Menell and Suzanne Scotchmer. Intellectual Property, 6 2005. forthcoming, Handbook of Law and Economics. Mitch Polinsky and Steven Shavell, eds. Amsterdam: Elsevier.

William Nordhaus. *Invention, Growth and Welfare: A Theoretical Treatment of Technological Change*. M.I.T. Press, 1969.

Lynne M Pepall and Daniel J Richards. Innovation, Imitation, and Social Welfare. *Southern Economic Journal*, 60(3):673–684, 1994. ISSN 00384038.

Rufus Pollock. Cumulative Innovation, Sampling and the Hold-up Problem. DRUID Working Papers 06-29, DRUID, Copenhagen Business School, Department of Industrial Economics and Strategy/Aalborg University, Department of Business Studies, 2006.

Frederic Scherer. Nordhaus' Theory of Optimal Patent Life: A Geometric Reinterpretation. *American Economic Review*, 62(3):422–427, 1972.

Suzanne Scotchmer and Jerry Green. Novelty and Disclosure in Patent Law. *The RAND Journal of Economics*, 21(1):131–146, 1990. ISSN 07416261.

David J. Teece. Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6):285–305, December 1986.

Michael Waterson. The Economics of Product Patents. *American Economic Review*, 80(4):860–69, September 1990.

Chapter 5

Forever Minus a Day? Some Theory and Empirics of Optimal Copyright

5.1 Introduction

The optimal level of copyright, and in particular, copyright term have been matters of some importance to policymakers over the last decade. For example, in 1998 the United States extended the length of copyright from life plus 50 to life plus 70 years, applying this extension equally to existing and future work.[1] More recently in the EU generally, and particularly in the UK, there has been an extensive debate over whether to extend the term of copyright in sound recordings.

Using a parsimonious framework based on those already in the literature (see e.g. Landes and Posner (1989); Watt (2000))[2] we analyze various questions related to the optimal level of copyright protection, deriving, under a simple set of assumptions, several novel results. In particular, we show that (a) optimal protection is likely to decrease as the cost of production for 'originals' falls (and vice-versa); (b) technological change which reduces costs of production may imply a decrease or an increase in optimal levels of protection (this contrasts with a large number of commentators particularly in the copyright industries who have argued that such change necessitates increases in protection); and (c) the optimal level of copyright will, in general, fall over time as the stock of work increases.

Note that costs are usually divided into those related to 'production', 'reproduction' and 'distribution' with the distinction between the first two being that production costs are those relating to the creation of the first instance of a work while reproduction relates to the costs of producing subsequent copies. However in this particular case we take 'production' costs to include all expenditures, fixed as well as variable, related to the creation *and* distribution of the first version of the work and all authorised reproductions thereof (these are often termed 'originals' in the literature in opposition to 'copies': unauthorised – though not necessarily illegal – reproductions of the work

[1]It was in a congressional speech prior to the enactment of the Copyright Term Extension Act (CTEA) that Mary Bono, widow of the musician Sonny Bono, famously referred to the proposal of Jack Valenti, president of the Motion Picture Association of America, to have copyright last for 'Forever minus a day': "Actually, Sonny wanted the term of copyright protection to last forever. I am informed by staff that such a change would violate the Constitution. . . . As you know, there is also Jack Valenti's proposal for the term to last forever less one day. Perhaps the Committee may look at that next Congress." (CR.144.H9952)

[2]There are, of course, analogies between the optimal patent literature commencing with Nordhaus (1969) and the optimal copyright literature. However the differences are such that the two areas remain largely distinct. Specifically because, crudely, patents are for 'ideas' while copyright is for 'expression' the issues of reuse and breadth while central to patent questions are much less important to copyright ones – similarly while in the patent literature it makes sense to consider several agents 'racing' for a specific innovation this has little meaning in copyright where works are so diverse and no two individuals are likely to produce something so directly substitutable. Conversely reproduction (the making of the 'copy') is a major factor in the analysis of copyright but is essentially irrelevant in the consideration of patents.

in question).

This first result is of particular interest because recent years have witnessed a dramatic, and permanent fall, in the costs of production of almost all types of copyrightable subject matter as a result of rapid technological advance in ICT and related fields. With the growth of the Internet costs of distribution have plummeted and will continue to do so as both the capacity and the level of uptake continue to increase. Similarly, cheaper computers, cameras, and software have had a significant impact on basic production costs in both the low and high end market.

One caveat needs to be mentioned here. As discussed, there is a distinction to be drawn both between authorised and unauthorised reproduction. The move to a digital environment reduces the costs of both of these types of activities – formally, there is a high degree of correlation between the changes in the costs of producing 'originals' and 'copies'. As a variety of authors have pointed out, a reduction in the cost of making 'copies', that is in the cost of unauthorised reproduction, may or may not necessitate an increase in the optimal level of protection – see e.g. (Johnson, 1985), (Novos and Waldman, 1984), Liebowitz (1985) and Peitz and Waelbroeck (2006). This impact of technological change on 'copies' as well as 'originals' is incorporated in our second result which shows that, when both effects are taken into account, the overall implications for the optimal level of protection are ambiguous. While such a result cannot give immediate guidance to policymakers, it does suggest one should be cautious about drawing 'obvious' conclusions about the implications of a digital environment for the level of copyright protection.

The third main result, that optimal protection falls over time, also has importance for policy. In most systems of law, it is extremely difficult to remove or diminish rights once they have been granted. Thus, once a given level of protection has been awarded it will be all but impossible to reduce it. However, according to our result, the optimal level of protection will decline over time (as the amount of work available grows). This being the case, a prudent policy-maker faced with uncertainty would want to be especially careful about increasing the level of copyright.

Finally, in the last section of the paper we turn to the specific case of copyright 'term' – that is, the duration of the copyright. Building on the framework already developed, we derive a single simple equation which defines optimal copyright term as a function of the key exogenous variables: the discount rate, the rate of 'cultural decay', the supply function for creative work and the associated welfare (and deadweight-loss) associated with new works. Combining this with empirical data we are able to provide one of the first theoretically *and* empirically grounded estimates of optimal copyright

term.[3]

5.2 A Brief Note on Copyright Law

The reader should be aware that the term of copyright varies both across jurisdictions and across types of protected subject matter. The right in a recording – as opposed to the underlying composition – is considered a 'neighbouring right' and is treated differently from a normal 'copyright'. In particular, signatories to the Berne convention (and its revisions) must provide for an 'authorial' copyright with a minimal term of life plus 50 years, recordings need only be protected for 50 years from the date of publication.

Furthermore, and rather confusingly, works can sometimes be moved from one category to the other as was the case with film in the UK following the implementation of the 1995 EU Directive on 'Harmonizing the Term of Copyright Protection' (which 'harmonised' copyright term up to life plus 70 years). Prior to this UK law had treated the copyright in the film itself as a neighbouring right and therefore accorded it a 50 year term of protection. Following the implementation of the Directive, the copyright in a film became an 'authorial' copyright and subject to a term of protection of life plus 70 years.[4]

5.3 Framework

In this section we introduce a minimal framework but one which is still rich enough to allow the derivation of our results.

The strength of copyright (also termed the level of protection) is represented by the continuous variable S with higher values implying stronger copyright. For our purposes here it will not matter exactly what S denotes but the reader might keep in mind, as examples, the length of copyright term and the breadth of the exclusions (conversely

[3]As Png (2006) notes, there is a lack of empirical work on copyright generally. Existing estimates of optimal term are very sparse. Boldrin and Levine (2005) calibrate a macro-oriented model and derive a figure of 7 years for optimal term in the United States. (Akerlof et al., 2002) in an examination of the US Copyright Term Extension Act argue, simply on the basis of the discount rate, that a term of life plus seventy years must be too long. By contrast, Liebowitz and Margolis (2005), argue that the current US term of life plus 70 years might not be too long – though they too do not provide an explicit model.

[4]That was not all, as Cornish and Llewelyn (2003, para. 10-45) note, 'the very considerable investment which goes into major film productions was held to justify a special way of measuring lives. To guard against the consequences of the director's early death, the longest life among "persons connected with the film" is taken; and these include not only the principal director but the author of the screenplay, the author of the dialogue and the composer of any specifically created film score.'

the narrowness of the exceptions from the monopoly right that copyright affords its owner).

Many possible works can be produced which may be labelled by 1,2,3, ... Let $N = N(S)$ denote the total number of works produced when the strength is S.[5] Note that N may also depend on other variables such as the cost of production, the level of demand etc. however we have omitted these variable from the functional form for the time being for the sake of simplicity.

Assumption 5.3.1. The form of the production function for copyrightable work.

1. At low levels of protection, increasing protection increases the production of works: $\lim_{S\to 0} N'(S) > 0$.[6]

2. Diminishing returns to protection: $N''(S) < 0$.

3. (optional) Beyond some level increasing protection further reduces production: $\lim_{S\to\infty} N'(S) < 0$.[7]

Each work created generates welfare for society, and we denote by w_i the welfare generated by the i'th work. The welfare deriving from a given work (once produced) depends on the strength of copyright, so $w_i = w_i(S)$ and it is assumed that increasing copyright reduces the welfare generated from a work so $w_i'(S) < 0$.

Total welfare, denoted by $W = W(N, S)$, is then the aggregation of the welfare from each individual work. This need not be a simple sum as we wish to allow for interactions between works – for example we would expect that as there are more and more works the value of new work declines. We shall discuss this further below, but for the time being we may leave the exact form of aggregation opaque.

Assumption 5.3.2. Using subscripts to indicate partial differentials:

1. Welfare is increasing in the number of works produced: $W_N > 0$.

[5]Throughout we shall gloss over the fact that N is discrete and allow the differential both of N and with respect to N to exist.

[6]Note that without this assumption we trivially have that the optimal level of copyright protection is zero. ('Proof': $\lim_{S\to 0} N'(S) \leq 0$, which, combined with the next assumption would imply $N'(S) \leq 0, \forall S \geq 0$. Thus production of works would be non-increasing in the level of protection. Combined with the fact that welfare per work is non-increasing in the level of protection (see below) this implies trivially that the optimal level of protection is zero – i.e. there should be no copyright). While there is nothing that a priori should exclude this possibility, as just shown, if it does not hold then the analysis is trivial. Thus rather than add to each statement of the results the rider that it depends on this assumption holding and if not then optimal copyright protection is zero we simply make the assumption here and those who are unconvinced of its validity should simply remember that this implies a zero level of protection.

[7]This assumption is based on a very similar one in Landes and Posner (1989). Unless otherwise stated this assumption will *not* be used when deriving any of the results below.

2. Keeping the number of works produced fixed, welfare is decreasing in the strength of copyright: $W_S < 0$ (this follows immediately from the assumption of diminishing welfare at the level of individual works).

3. Diminishing marginal welfare from new works: $W_{NN} < 0$.

Since the number of works produced is itself a function of the level of copyright we may eliminate N as an argument in W and write:

$$W = W(S) = W(N(S), S)$$

Where it is necessary to distinguish the different forms of the welfare function we shall denote this version as the 'reduced form'. Finally note that, assuming *only* that $lim_{S \Rightarrow \infty} W(S)$ exists (with the value of infinity permitted), then as $[0, \infty]$ is compact (using the circle projection) and $W(S)$ is a continuous function (in the induced topology), W has a unique maximum somewhere in this range. As this is the welfare maximizing level of protection we term this the *optimal* level.[8]

Finally before commencing on the derivation of results we require the technical assumption that all functions are continuous and at least twice continuously differentiable.

5.4 The Relation of the Production and Welfare Maximising Levels of Protection

Lemma 5.4.1. *Under assumptions 5.3.1.1 and 5.3.1.2 there exists a unique level of protection which maximizes the production of creative work. We denote this by S^p. Furthermore, EITHER there exists a finite solution to $N'(S) = 0$ and this is S^p OR no such solution exists and $S^p = \infty$. With assumption 5.3.1.3 only the first option is possible.*

Proof. By Assumption 5.3.1.1 N is increasing when the level of protection is 0 (the lowest possible) thus 0 cannot be a maximum. By Assumption 5.3.1.2 if a finite maximum exists it must be unique and this maximum must be a solution of $N'(S) = 0$ (if there is such a solution then N' is negative from that solution onwards so infinity is not a solution). If no such solution exists then for all $S > 0$ we have $N'(S) > 0$ and the maximizing level of protection is infinite. □

[8]Note that it is possible that there are multiple levels of protection which achieve the welfare maximum – for example consider the case of $W(S) =$ constant. In this case take as the *optimal* level the minimum (infinum) of these welfare maximizing levels of protection.

Theorem 5.4.2. *If the level of protection which maximizes the production of copyrightable work, S^p, is finite then the optimal level of protection, S^o, is strictly less than S^p.*

Proof. If S^p is finite then $N'(S^p) = 0$ and since $N''(S) < 0$ we have $N'(S) \leq 0, \forall S \geq S^p$. Marginal welfare is:

$$W'(S) = \frac{dW(S)}{dS} = \frac{dW(N(S), S)}{dS} = N_S W_N + W_S$$

Now $W_S < 0, \forall S$, so combining this with the properties of the work production function, $N(S)$, we have that:

$$\forall S \geq S^p, W'(S) < 0$$

Hence, welfare is already declining at S^p and continues to decline thereafter. Thus, the optimal, that is welfare maximizing, level of protection, S^o, must lie in the range $[0, S^p)$.

□

Remark 5.4.3. If the level of protection which maximizes the production of copyrightable work, S^p, is infinite then no immediate statement can be made as to whether the optimal level of protection, S^o, will be finite (and hence less than S^p) or infinite.[9]

From this point on we make the following assumption:

Assumption 5.4.4. The optimal level of protection is finite, and is the unique level of protection, S^o, satisfying $W'(S^o) = 0, W''(S^o) < 0$.

5.5 Production Costs and the Optimal Level of Protection

Let us now introduce production costs by writing $N = (S, C, U)$ where C is a variable denoting production costs of 'originals' (authorised reproductions) and U a variable denoting the production cost of 'copies' (unauthorised reproductions) (we do not need

[9]For example, consider a very simple multiplicative structure for total welfare of the form: $W(S) = f(N(S))w(S)$ with $f(N)$ any functional form with $f' > 0, f'' < 0$ (e.g. $N^a, a \in (0,1)$). Then taking any function $g(S)$ with $g' > 0, g'' < 0$ and defining $N(S) = g(S), w(S) = g(S)^{1-a+\epsilon}, \epsilon \in (0, a)$ we have a setup satisfying Assumptions 5.3.1 (excluding 5.3.1.3) and 5.3.2 and with $W(S) = g(S)^\epsilon$ – a welfare function whose maximising level of protection is clearly infinite.

Finally note that this does *not* require that the number of works produced be infinite, for example we could have $g(S) = 1 + K - K/(1 + S)$ in which case there is a finite upper bound on the number of works produced.

to be specific here as to their form so these may be marginal costs or fixed costs or both).[10] We assume that:

1. For any given level of protection, as the costs of 'originals' increase (decrease) production decreases (increases): $N_C < 0$. This follows from the fact that increases in cost reduce profits (revenues are constant).

2. For any given level of protection, as the costs of 'copies' (unauthorised reproductions) increase (decrease) production increases (decreases): $N_U > 0$ (NB: the costs of 'originals' are assumed to remain unchanged). The reasoning behind this is that 'copies' compete with 'originals' and hence increases (decreases) in the cost of 'copies' raise the revenues to 'originals' (more formally, as 'copies' and 'originals' are substitutes so the cross price derivatives of demand are negative). This in turn raises profits to the owners of 'originals' and hence increases production.[11]

We also need to take account of the impact of costs on welfare. To reflect this we rewrite welfare as a function of both the level of protection *and* the level of costs: $W = W(S, C, U) = W(S, C, U, N(S, C, U))$.[12]

Lemma 5.5.1. *Take any exogenous variable X which affects the welfare function (whether directly and/or via its effect on production N). Assuming that the initial optimal level of protection, S^o, is finite, if $d^2W(S^o)/dXdS$ is positive then an increase (decrease) in the variable X implies an increase (decrease) in the optimal level of protection.*

[10]Note that we would usually assume that the cost of making 'copies' is itself, at least partially, a function of the level of protection. However here we prefer to keep the effect of the level of protection and of the cost of making 'copies' distinct. Thus, it is perhaps better to think of U as encapsulating copying costs as determined purely by exogenous factors such as technology.

[11]The assumption that decreases in the cost of unauthorised copying are unambiguously bad for the producers of copyrightable works is a standard one. However, there are at least two factors which operate in the opposite direction. First, 'copiers' still need to purchase 'originals' and thus producers of 'originals' may still be able to extract rents from 'copiers' by raising the price of originals much in the way that the price of a first-hand car takes account of its resale value on the second-hand market (see Liebowitz (1985)). Second, greater dissemination of a work due to unauthorized copying may lead to increase in demand for 'originals' or for complementary goods, particularly if 'copies' and 'originals' are not perfect substitutes. For a recent theoretical model see Peitz and Waelbroeck (2006). Empirical work, mainly centred on the impact of unauthorised file-sharing on music sales has, as yet, provided no decisive answer as to whether 'sampling' may outweigh 'substitution' (see, for example, the contradictory results of Oberholzer and Strumpf (2007) and Blackburn (2004)). Given these uncertainties, we feel it prudent to stick with the straightforward, and conservative, assumption that decreases in the cost of unauthorised copying decreases the production of creative work.

[12]Note here that total welfare depends both directly and on costs and indirectly via production. We have just discussed the indirect impact via production and we discuss the direct impact further below when signing the partial derivatives of W however it should be obvious that there is a direct impact of costs on welfare because higher (lower) production costs, whether of originals or copies, reduces (increases) producer surplus for a given work and hence reduces (increases) welfare for that work.

Proof. Denote the initial optimum level of protection, where X is at its initial value, by S^o. Since we are a finite optimum we have that at S^o:

$$W'(S^o) = N_S W_N + W_S = 0 \tag{5.1}$$
$$W''(S^o) < 0 \tag{5.2}$$

Suppose, X now increases. Since $d^2W/dXdS$ is positive we must now have: $W'(S^o) > 0$. For small changes in X, $W''(S^o)$ is still negative and thus protection must increase to some $S^{o2} > S^o$ in order to have $W'(S^{2o}) = 0$; and S^{2o} is the new optimum level of protection. □

5.5.1 Production Costs

Let us consider first, what occurs is there is an increase (or conversely a decrease) in the costs of producing 'originals' with all other exogenous variables, including the cost of producing 'copies', unchanged. Substituting C for X we have:

Corollary 5.5.2. *If $d^2W(S^o)/dCdS > 0$ then an increase (decrease) in costs of 'originals' implies an increase (decrease) in the optimal level of protection.*

Given the importance of signing $d^2W/dCdS$ let us explore further by working through the differential:

$$\frac{d^2W}{dCdS} = \frac{d}{dC}(N_S W_N + W_S) = N_{CS} W_N + N_S W_{NN} N_C + N_S W_{CN} + W_{SN} N_C + W_{CS}$$

Now:

1. $W_C < 0$ – welfare declines as costs rise because higher costs for a given work mean less producer welfare, and hence less total welfare, from that work.

2. $W_{NS} < 0$ – increasing S for a given work reduces welfare (which is why $W_S < 0$) and thus increasing the number of works increases the negative effect on total welfare.

3. $W_{CS} \geq 0$ – the marginal effect of increasing protection declines as costs rise (remember W_S is negative).

4. $W_{CN} \leq 0$ – increasing production costs reduces the marginal benefit of new work (as each new work provides less welfare).

5. $N_{CS} > 0$ – the marginal impact of protection declines with lower costs.

The last inequality is the least self-evident of these. One justification for it is as follows: the level of production is a function of the level of (average) profit, π, per work: $N = g(\pi)$. With diminishing returns we would expect $g'' < 0$. Profits can be broken up into income and costs, $\pi = I - C$, with the level of protection *only* affecting income and not costs. In that case we have $N_{CS} = g''\pi_S\pi_C > 0$

Furthermore, by prior assumption or analysis we have: $W_N > 0, N_S > 0, W_{NN} < 0, N_C < 0, W_S < 0$. Thus, four of the five terms in the equation for the mixed second-order derivative for welfare are positive while one, $N_S W_{CN}$ is not.

This means, that we cannot unambiguously say whether an increase or decrease in the costs of 'originals' implies an increase or decrease in the level of protection. In some ways this is somewhat surprising. Increased costs reduces the number of works and reduces the deadweight loss per work from protection so we might expect that increasing protection would unambiguously improve welfare.

The reason this is not necessarily so is that increased costs also reduce the welfare per work and hence while the number of works falls, which increases the marginal value of a new work, the increase in costs provides a countervailing effect (W_{CN}). As a result it is possible that the reduction in welfare per work due to higher costs is so dramatic as to outweigh all the other effects which favour an increase in term. Thus, a general statement based on theory alone is not possible.

That said, all of the reduction in welfare comes via a reduction in producer surplus due to higher costs. Hence the proportional reduction in income, and hence output, is likely to be substantially higher than the proportional reduction in welfare. As a result one would expect the effect of a reduction in output (N) to outweigh the effect of a reduction in welfare and therefore for $d^2W/dCdS$ to be negative. Formalizing this condition we have:

Proposition 5.5.3. *Assuming an initial finite optimal level of copyright, a sufficient condition for a reduction in the cost of 'originals' (leaving other variables unchanged) to imply a reduction in the strength of copyright is that an increase in costs C, results in an increase in the marginal value of new work: $\frac{d}{dC}W_N > 0$.*

5.5.2 Technological Change

Let us now introduce 'technological' change explicitly as a variable T. We shall assume that T has no direct effect on welfare but only operates through its impact on the costs of 'originals' and 'copies' (C and U), and does so by reducing both types of costs (so

$C_T < 0, U_T < 0$). Thus total welfare now has the form $W(S,T) = W(S,C(T),U(T)) = W(S,C(T),U(T),N(S,C(T),U(T)))$. Substituting T for X in Lemma 5.5.1 we have:

Corollary 5.5.4. *If, at the current optimal level of protection, $d^2W/dTdS < 0$ then technological change implies a reduction in the level of copyright. Conversely if $d^2W/dTdS > 0$ then an increase in the level of copyright is required.*

Turning again to an explicit consideration of the second derivative we have:

$$\frac{dW^2}{dTdS} = \frac{d}{dT}(N_S W_N + W_S) = N_{TS}W_N + N_S W_{NN} N_T + N_S W_{TN} + W_{NS} N_T + W_{TS}$$

Focusing on the effect on the output of works: $N_T = N_C C_T + N_U U_T$, the effect of technological change will be ambiguous: the first term is positive since improvements in technology reduce the costs of originals ($C_T < 0$), while the second is negative since production goes up (down) as the cost of unauthorised copying decreases (increases): $N_U > 0$. However unlike welfare, N is (easily) observable, and it seems that recent years have seen an increase in the amount of work available. Thus, let us assume $N_T > 0$. We then have:

1. $N_{TS} < 0$ – as costs drop value of increasing protection diminishes (as the number of works is increasing).

2. $W_{TN} > 0$ – marginal value of new work increases as T increases (a reduction in both types of costs increases welfare: $W_U, W_C < 0$).

3. $W_{NS} < 0$ – see above.

4. W_{TS} is ambiguous – increasing T reduces both C and U and while a reduction in the costs of 'originals' increases deadweight losses a reduction in U reduces them with the overall effect ambiguous.

Thus, we have:

$$\frac{dW^2}{dTdS} = \text{-ve} + \text{-ve} + \text{+ve} + \text{-ve} + ?$$

In many ways this is similar to the previous situation. However the ambiguities here are more pronounced. In particular, one term can not be signed unambiguously from theory alone (W_{TS}) and it is less likely that the 'contrary' term here, $N_S W_{TN}$, will be small relative to the others. The key trade-off then is similar to the one discussed above.

On the one hand technological change reduces costs and thereby increases output which diminishes the value of new work (implying a reduction in copyright). However, at the same time, by reducing costs technological change increases the value of new work. These two effects operate in opposite directions and it is not a priori clear which will be the stronger. Again one might argue that the proportional increase in incomes for producers is likely to be at least as large as the increase in welfare and hence the increase in output will more than offset the impact on welfare per work. However, one must be cautious here because technological change may also reduce deadweight losses via a reduction in the cost of unauthorised copying and overall it would seem impossible to draw unambiguous conclusions from theory alone.

5.5.3 Discussion

Examples of cost-reducing technological change are ubiquitous in recent years arising, in the main, from the move to a digital environment. As discussed, focusing on the case of 'originals' alone, it seems likely that such changes would imply a reduction in the optimal level of copyright. However, this gives only half the story – technological change is likely to reduce the costs of both 'originals' and 'copies' at the same time. While it is unclear whether technological advance has reduced the costs of one faster than the other – the reductions in both cases seem dramatic – it appears that the overall level of output has risen. Using this fact, we examined whether optimal term should rise or fall as technological progress reduces costs. While based on theory alone, it was not possible for an unambiguous answer to be given, we were able to characterise (and sign) most of the main factors impacting on welfare.

This ambiguous result is not surprising given the contrary effects at play. Furthermore our work highlights the key terms in need of empirical estimation in order to obtain an unambiguous conclusion regarding the implications of technological change for copyright.[13] We also think it important in demonstrating that care must be taken when drawing 'obvious' conclusions for copyright policy from changes in the external environment. Much of the motivation for strengthening copyright in recent years, whether by extending term or by the addition of legal support for technological protection measures (TPMs) – as in the WIPO Copyright Treaty of 1996 and its subsequent translation into national laws such as the DMCA (1998) and the EUCD (2001) – has been based on the implicit assumption that the move to a digital environment necessitated an increase in the strength of copyright because technological change made

[13]This is very similar to situation regarding copyright term which we address below. There too theory cannot tell us what level of term is optimal but can help us pinpoint the key variables in need of empirical estimation.

unauthorised copying ('piracy') easier. But focusing only on the reduction in the costs of unauthorised copies ignores the impact of technology on authorised production and distribution. As we have shown, such an approach omits a major part of the overall picture and may lead to erroneous conclusions regarding both the necessity and direction of policy changes.

5.6 Optimal Copyright in a Dynamic Setting

Our previous analysis has dealt only with a static setting in which all production could be aggregated into a single figure, N. In this section we will need to enrich this basic approach by introducing 'time'. To do this let us define n_t as the number of works produced in time period t and N_t as the number of works available to society in period t.[14] N_t will be the 'real' or 'effective' amount of work available, that is it takes account of cultural depreciation and obsolescence – which represent the fact that many works are 'of their time' and are, or at least appear to be, of little value to future generations. Specifically we expect N_t not to be the absolute amount of past and present work available but rather an 'equivalent' amount denominated in the same terms as n_t. Formally, if we let $b(i)$ be the 'rate of cultural decay' after i time periods ($b(0) = 1$), then the 'effective' amount of work in period T is the sum of the production of all previous periods appropriately weighted by the level of cultural decay:

$$N_t = \sum_{i=0}^{\infty} b(i) n_{t-i}$$

Then, defining $d(i)$ as the discount factor to period i, the total welfare calculated at time t is:

$$W_t^{Tot}(S) = \sum_{i=0}^{\infty} d(i) W(N_{t+i}(S), S)$$

We shall assume this is single-peaked and differentiable (so the first-order condition is necessary and sufficient).[15]

Theorem 5.6.1. *Assume that at time $t = 0$ production is approximately zero (this could be for several reasons the most obvious being that this type of work only comes*

[14]Both numbers will have the same set of arguments as the static N we had before so we will have $n_t = n_t(S, C)$, $N_t = N_t(S, C)$ though note that if the arguments can vary over time then the arguments would have be modified appropriately (those to n would need to include future values and those for N both past and future values).

[15]This dynamic problem has substantial similarities with the standard optimal control problems of dynamic growth models. Specifically, let $b(i)$ takes a standard exponential form $b(i) = \beta^i$ and allow

into existence at this point[16]*). Then, assuming that sequence of works produced per year,* n_i *is such that* $N(t) = \sum_{i=0}^{t} b(t-i)n_i$ *is non-decreasing, optimal protection declines over time asymtoting towards what we term the 'steady-state' level.*

Proof. We first provide an informal justification for this result before turning to a formal, mathematical, 'proof'.

No works are produced before time zero so, as time increases, the backlog of work will grow. As the backlog grows a) the value of producing new work falls and b) the welfare losses from increased protection are levied not just on new works but on the backlog as well.

To illustrate consider the situation with respect to books, music, or film. Today, a man could spend a lifetime simply reading the greats of the nineteenth century, watching the classic movies of Hollywood's (and Europe's) golden age or listening to music recorded before 1965. This does not mean new work isn't valuable but it surely means it is less valuable from a welfare point of view than it was when these media had first sprung into existence. Furthermore, if we increase protection we not only restrict access to works of the future but also to those of the past.

As a result the optimal level of protection must be lower than it was initially in fact it must fall gradually over time as our store of the creative work of past generations gradually accumulates to its long-term level. We now turn to the formal argument.

S to be set anew each time period (it can then take the role of a standard control variable). Then:

$$\begin{aligned} N_t &= \beta N_{t-1} + n_t \\ n_t &= f(S_t, S_{t+1}, ..., N_t, N_{t+1}, ...) \\ W_t &= W(N_t, S_t) \\ W_t^{Tot} &= W_t + \beta \sum_{i=0}^{\infty} \beta^i W_{t+1+i} = W_t + \beta W_{t+1}^{Tot} \end{aligned}$$

Then, comparing to growth models, N_t is K_t (capital), n_t is Y_t (production), S_t is c_t (the control variable – usually consumption), W_t is $U(c_t)$ (utility from consumption) and W_t^{Tot} is the value function (overall welfare). Of course our setup is more complex than the standard growth framework since output (the number for works produced) depends not just on current values for the control variable but on future values of the control variable and future levels of output (this is because creative works are durable).

We note that these sorts of problems have been extensively analyzed – see Stokey, Robert E., and Prescott (1989) for a mathematical survey – and while it is relatively straightforward to ensure the existence of an equilibrium it is hard to state any general results about the time paths of the state and control variables (see e.g. the 'anything goes' result of Boldrin and Montrucchio (Stokey, Robert E., and Prescott, 1989, Thm 6.1) which demonstrates that any twice-differentiable function g can be obtained as the policy function of a particular optimal dynamic growth problem).

Thus, here we restrict to the case where the control variable may only be set once (S is given forever) and we also assume, when stating our result, that the time path of the number of works ('capital') is non-decreasing – a result obtained in many, though not all, growth models and which, in the case of copyright, appears to fit well with the available data.

[16]For example, films only came into existence around 1900, and sound recordings were only possible from the late 19th century.

Optimal protection, S^t, at time t solves:

$$\max_S W_t^{Tot}(S)$$

The first-order condition is:

$$\frac{dW_t^{Tot}(S^t)}{dS} = 0$$

Consider this at time t then:

$$\sum_{i=0}^{\infty} d(i) \frac{dW(N_{i+t}(S^t), S^t)}{dS} = 0$$

Recall that $\frac{\partial}{\partial N}\frac{dW}{dS} < 0$ (the marginal value of protection goes down as the number of works increases and the total deadweight loss increases) so that, if $N^1 > N^2$:

$$\frac{dW(N^1, S)}{dS} < \frac{dW(N^2, S)}{dS}$$

Now, by assumption on the structure of n_i, $\forall i, N_{i+t+1} > N_{i+t}$. Thus, we must have:

$$\frac{dW_{t+1}^{Tot}(S^t)}{dS} = \sum_{i=0}^{\infty} d(i) \frac{dW(N_{i+t+1}(S^t), S^t)}{dS} < \sum_{i=0}^{\infty} d(i) \frac{dW(N_{i+t}(S^t), S^t)}{dS} = \frac{dW_t^{Tot}(S^t)}{dS} = 0$$

So we have that:

$$\frac{dW_{t+1}^{Tot}(S^t)}{dS} < 0$$

Since W^{Tot} is single-peaked this implies that the level of protection which maximizes W_{t+1}^{Tot} must be smaller than S^t. That is the optimal level of protection at $t+1, S^{t+1}$, is lower than the optimal level of protection at t, S^t.

Finally, we show that the optimal level of protection will tend to what we term the steady-state level. We have just proved that S^t is a declining sequence. Since values for S are bounded below by 0 by Bolzano-Weierstrass we immediately have that the sequence must converge to a unique $S = S^\infty$. By analogous arguments associated with this 'steady-state' level of protection will be a steady-state level of output per period n^∞ and effective number of works N^∞.

□

5.6.1 Remarks

The preceding result has important implications for policy. In most systems of law, it is extremely difficult to remove or diminish rights once they have been granted. Thus, in most circumstances, once a given level of protection has been granted it will be all but impossible to reduce it. However, according to the preceding result, in general the optimal level of protection will decline over time.

In many ways this is a classic 'dynamic inconsistency' result: the preferences of a welfare-maximizing policy-maker at time zero are different from those at some future point T.[17] Furthermore, it is clear that no particular point in time has any more validity over any other point as regards being chosen as a reference point. Moreover, from the perspective of any given point in time the ability to 'commit' to a given level of protection may be very valuable.[18] That said the result is still important for two reasons.

First, whether because of a paucity of data or disagreement about the form of the model, there is frequently significant uncertainty about the optimal level of protection. But one thing we do know from the preceding result is that, whatever optimal level of protection currently, it will be lower in the future. Combined with the asymmetry in decision-making already mentioned – namely, that it is much harder to reduce protection than to extend it – this implies it is prudent for policy-makers to err on the low side rather than the high side when setting the strength of copyright.

Second, and more significantly this result provokes the question: if optimal protection should decline over time why does the history of copyright consists almost entirely of the opposite, that is to say, repeated increases in the level of protection over time (duration, for example, has been increased substantially in most jurisdictions since copyright was first introduced[19]). After all, while one can argue that for 'commitment' reasons a policy-maker would not reduce the level of protection over time, our result certainly runs counter to the repeated increases in protection, many of which have taken place in recent years (when the stock of copyrightable works was already large).

The obvious answer to this conundrum is that the level of protection is not usually determined by a benevolent and rational policy-maker but rather by lobbying. This

[17] However we should note some important differences. In the classic case of dynamic inconsistency, even at stage two (in a two-stage game) the policy-maker would have preferred to have been able to commit at stage one (to a different policy). By contrast here the policy-maker at stage two simply has a different optimum policy than at stage one – i.e. the stage two policy (which includes specifying action at at all stages including previous ones) is optimal from the point of view of stage two but is not optimal from the point of view of stage one (and vice-versa).

[18] It is precisely concerns over the ability of a policy-maker to credibly commit to a particular macroeconomic target that animates many of the traditional models of dynamic inconsistency.

[19] Most prominently in recent times in the United States in 1998 and in the EU in 1995.

results in policy being set to favour those able to lobby effectively – usually groups who are actual, or prospective, owners of a substantial set of valuable copyrights – rather than to produce any level of protection that would be optimal for society as a whole. Furthermore, on this logic, extensions will be obtained precisely when copyright in existing, and valuable, material is about to expire. In this regard it is interesting to recall that many forms of copyrightable subject matter are of relatively recent origin. For example, the film and recording industry are only just over a hundred years old with the majority of material, in both cases, produced within the last fifty years. In such circumstances, and with copyright terms around 50 years, it perhaps not surprising that the last decade has seen such a flurry of extensions and associated rent-seeking activities.

5.7 Optimal Copyright Term

We now turn to the case of optimal copyright term. By interpreting the level of protection, S, as the length of copyright the framework set out above can be re-applied directly. At the same time, because we are now dealing with a more specific case we can add greater structure to the model and, by so doing, obtain some sharper predictions. Our aim here is to derive a numerical, quantitative estimate, for the length of copyright term. Clearly this will be an empirical task and the main use of theory in this section will be in characterising optimal term as a function of underlying variables that can be feasibly estimated from available data.

As the reader will recall, the basic trade-off inherent in copyright is between increasing protection to promote the creation of more work and reducing protection so as to gain more from existing work. The question of term, that is the length of protection, presents these two countervailing forces particularly starkly. By extending the term of protection the owners of copyrights receive revenue for a little longer. Anticipating this, creators of work which were nearly, but not quite, profitable under the existing term will now produce work, and this work will generate welfare both current and future welfare for society. At the same time, the increase in term applies to all existing works – those which would have been created under the initial level of copyright. Since extending term on these works prolongs the copyright monopoly it reduces total welfare as a result of the extra deadweight loss.

It is these two, contrary, effects which will form the main focus of our investigation here. Together they will already provide with plentiful matter for theoretical and empirical efforts but we should note that in confining ourselves in this we we will be ignoring a variety of further issues. For example, much creative endeavour builds upon

the past and an extension of term may make it more difficult or costly do so – were Shakespeare's work still in copyright today it is likely that this would substantially restrict the widespread adaptation and reuse that currently occurs. However we make no effort to incorporate this into our analysis despite its undoubted importance (it is simply too intractable from a theoretical and empirical perspective to be usefully addressed at present). We will also ignore questions of ex post investment, that is investment by a copyright owner after creation of the work, as well as inefficient exploitation, that is a failure by a copyright owner to maximize the value of the work in their possession.[20]

5.7.1 Theory

Our first step then is to link the two main effects to a common set of underlying variables. We begin by introducing explicit consideration of the revenue from a work. We shall assume that without copyright revenue is zero. Let revenue (under copyright) on the jth work in the ith period after a work's creation be given by $r_j(i)$ and present value of total revenue to period T be $R_j(T)$ (where implicitly we assume that T is less than the current copyright term). Let $d(i)$ be, as above, the discount factor up to time i, then:

$$R_j(T) = \sum_{t=0}^{t=T} d(t) r_j(t)$$

Revenue decays over time due to 'cultural decay'. We specify cultural decay by $b(t)$, with $r_j(t) = b(t)r(0)$ (cultural decay is assumed to occur at the same rate independent of the work), so that we have:

$$R_j(T) = \sum_{t=0}^{t=T} d(t) b(t) r_j(0)$$

As was shown above, the level of optimal copyright will not be constant over time even if all underlying parameters stay constant. This variation is not our focus here. Instead we are interested in how the basic parameters – the discount rate, the level of cultural decay etc – affect the optimal level of copyright. Thus, when comparing two terms here we shall compare them at their long-run, steady-state, level.[21] Formally,

[20] See e.g. Landes and Posner (2003) on ex post investment and Brooks (2005) for evidence on inefficient exploitation. We should note that, in our opinion, both of these effects are likely to be relatively limited, and hence we believe their omission, unlike that of 'reuse', is unlikely to have a serious impact on the overall results.

[21] It it is in this assumption that we differ most significantly from previous analyses such as that of Landes and Posner (1989). Their model implicitly assumes no work already exists and therefore,

the following assumptions will be made in what follows:

1. All calculations will be of a comparative static nature with the level of production taken at its long run equilibrium value. Thus we take the amount of work produced per period n_t to be the constant and equal to the steady-state level which we will denote by n. Similarly the 'effective' amount of work available per period will be constant and will denote it by N.

2. Discount factors are the same for producers and for society (i.e. we discount welfare at the same rate we discount income for producers).

3. Revenue and welfare (and dead-weight loss) per work experience the same rate of cultural decay. Thus total welfare per period may be obtained by summing over all vintages of works weighted by the relevant cultural decay.

Since we evaluate welfare at the long run equilibrium, production per period and welfare per period may be taken to be constant and equal to their long run equilibrium values. Therefore in what follows we focus on welfare per period (converting to total welfare is a trivial matter). We have the following result:

Theorem 5.7.1. *The marginal change in (per period) welfare with respect to an increase in the term of protection, S, when the current term is S^1, is as follows:*

$$\frac{dW(S^1)}{dS} = ns(n)y(n)b(S^1)\left(d(S^1)\frac{\sum_{i=0}^{\infty} b(i)}{\sum_{i=0}^{\infty} d(i)b(i)} + d(S^1)\frac{z(n)}{y(n)}\frac{\sum_{i=S^1}^{\infty} b(i)}{\sum_{i=0}^{\infty} b(i)d(i)} - \theta(n)\right)$$

Where:

$$\begin{aligned}
d(t) &= \textit{Discount factor to time } t \\
b(t) &= \textit{Cultural decay to period } t \\
y(j) &= \textit{Welfare (under copyright) from an extra } j\textit{th} \text{ new } \textit{work} \\
z(j) &= \textit{Deadweight-loss under copyright on the } j\textit{th} \text{ new } \textit{work} \\
\overline{z}(n) &= \textit{Average deadweight-loss under copyright on works } 1 \cdots n \\
s(n) &= \textit{Elasticity of supply of works with respect to revenue when there are } n \textit{ works} \\
\theta(n) &= \textit{Ratio of avg. d/w loss to welfare from new works} = \frac{\overline{z}(n)}{s(n)y(n)}
\end{aligned}$$

in the formulation of the previous section, maximizes welfare from the perspective of a social planner at time $t = 0$ rather than at the steady-state. We believe that the steady-state analysis presented here, which includes the prospective and retrospective effects of changes in copyright term, is the more appropriate – particularly since today most forms of copyrightable work have been produced for decades if not centuries.

In particular define the 'determinant', Δ as the bracketed term above, i.e.

$$\Delta = d(S^1)\frac{\sum_{i=0}^{\infty} b(i)}{\sum_{i=0}^{\infty} d(i)b(i)} + d(S^1)\frac{z(n)}{y(n)}\frac{\sum_{i=S^1}^{\infty} b(i)}{\sum_{i=0}^{\infty} b(i)d(i)} - \theta(n)$$

Then the optimal copyright term is determined by reference to the 'determinant' alone and is the solution of $\Delta = 0$.

Proof. As we are going to take derivatives we shall take all necessary variables (number to works, time etc) to be continuous rather than discrete (and we therefore have integrals rather than sums). Note that the conversion back to the discrete version is straightforward (but would make the notation and proof substantially more cumbersome).

We can express welfare per period as:

$$\begin{aligned} W &= \text{Welfare under infinite copyright} + \text{Extra welfare on works out of copyright} \\ &= \int_{j=0}^{n(S)} \int_{i=0}^{\infty} y(j)b(i) + \int_{j=0}^{n(S)} \sum_{i=S}^{\infty} z(j)b(i) \end{aligned}$$

With the first sum being over the n works produced each period and the second being over past periods ($i = 1$ corresponding to the period previous to this one, $i = 2$ to two periods ago etc). With the double sum we cover all works ever produced, bringing them up to the present in welfare terms by multiplying by a suitable amount of 'cultural decay'.

Differentiating we have:

$$\begin{aligned} \frac{dW}{dS} &= n'y(n)\textstyle\int_{i=0}^{\infty} b(i) + n'z(n)\int_{i=S}^{\infty} b(i) \quad -b(S)\displaystyle\int_{j=0}^{n} z(j) \qquad (5.3) \\ &= \text{Gain in welfare from new works} \quad -\text{Extra deadweight loss on existing works} \end{aligned}$$

Let us re-express the increase in the number of works, $n'(S^1)$, in terms of the change in revenue for a marginal (nth) work R_n (note we suppress the n subscript for terminological simplicity):

$$n' = \frac{dn}{dS} = n\frac{n'(R(S))}{n} = n\frac{dn}{dR}\frac{R}{n}\frac{R'}{R}$$

The middle term of the final expression is the elasticity of supply with respect to revenue, $s(n)$, while the last is the percentage increase in revenue. Total revenue on the marginal work itself equals (remember that once out of copyright revenue per period is zero):

$$R(S) = \int_{i=0}^{S} d(i)b(i)r(0) \implies R'(S) = d(S)b(S)r(0)$$

Thus, substituting, using $\overline{z}$, for average z and converting back to summations we have:

$$\begin{aligned}\frac{dW(S^1)}{dS} &= ny(n)d(S^1)b(S^1)r(0)\left(\frac{\sum_{i=0}^{\infty} b(i)}{\sum_{i=0}^{\infty} d(i)b(i)r(0)} + \frac{z(n)}{y(n)}\frac{\sum_{i=S^1}^{\infty} b(i)}{\sum_{i=0}^{\infty} b(i)d(i)r(0)}\right) - b(S^1)n\overline{z}(n) \\ &= ns(n)y(n)b(S^1)\left(d(S^1)\frac{\sum_{i=0}^{\infty} b(i)}{\sum_{i=0}^{\infty} d(i)b(i)} + d(S^1)\frac{z(n)}{y(n)}\frac{\sum_{i=S^1}^{\infty} b(i)}{\sum_{i=0}^{\infty} b(i)d(i)} - \frac{\overline{z}(n)}{s(n)y(n)}\right)\end{aligned}$$

□

Having now obtained an expression which characterises the optimal copyright term (Δ) our next task is to obtain estimates for its various component variables (b, d, θ etc). We go through each of the variables in turn, starting with the simplest to estimate (the discount rate) and progressing to the hardest (θ).

5.7.2 The Discount Rate

We assume a standard geometric/exponential form for the discount function. The relevant discount factor to use here is that related to those producing works so a plausible range is a discount rate in the range 4-9%. For example, CIPIL (2006) in considering a similar issue report that: Akerlof et al. (2002) use a real discount rate of 7%, Liebowitz in his submission to the Gowers review on behalf of the IFPI (International Federation of the Phonographic Industry) uses a figure of 5%, while PwC's report to the same review on behalf of the BPI (British Phonographic Industry) use the figure of 9%. Where we need to use a single value we will by default use a rate or 6% (corresponding to a discount factor of 0.943).

5.7.3 The Rate of Cultural Decay

We assume an exponential form for the cultural decay so that $b(i) = b(0)^i$ with $b(0)$ the cultural decay factor.[22] A plausible range for this cultural decay rate is 2-9% and by default we will use 5% (corresponding to a factor of 0.952). Since values for these

[22]It is likely that an exponential distribution is not a perfect fit for the cultural decay rate. In general, it appears that the rate of decay is sharper than an exponential for young works but flatter than an exponential for old works. This suggests that hyperbolic cultural decay might be a better model (just as hyperbolic discounting may be more accurate than exponential discounting for income). However, an exponential form appears to be a reasonable approximation and it is substantially more tractable. Thus we retain it here rather than using the more complex hyperbolic approach (just as an exponential form is regularly used for time discounting for analogous reasons).

variables are less well-established than those for the discount rate the evidence on which they are based merits discussion.

The prime source is CIPIL (2006), which reports estimates made by PwC based on data provided by the British music industry which indicate decay rates in the region of 3-10%. As these come from the music industry itself, albeit indirectly, these have substantial authority. To check these we have performed our own calculations using data on the UK music and book industry and obtain estimates for the rate of decay that are similar (in the case of music) or even higher (in the case of books).

Evidence from elsewhere includes the Congressional Research Service report prepared in relation to the CTEA (Rappaport, 1998). This estimates projected revenue from works whose copyright was soon to expire (so works from the 1920s to the 1940s). Rappaport estimates (p.6) that only 1% of books ever had their copyright renewed and of those that had their copyright renewed during 1951 to 1970 around 11.9% were still in print in the late 1990s. The annual royalty value of books go from $46 million (books from 1922-1926) to $74 million (books from 1937-1941). Turning to music, Rappaport focuses on songs (early recordings themselves have little value because of improvements in technology) and finds that 11.3% of the sample is still available in 1995. Annual royalty income rises from $3.4 million for works from 1922-1926 to $15.2 million for works from 1938-1941.

These figures correspond, in turn, to cultural decay rates of 3.2% and 10.5% respectively. However these are far from perfect estimates since we only have two time points. Furthermore these time points correspond to different 'cohorts' of work – which makes it difficult to disentangle decay effects from cohort effects, and both these cohorts are of fairly old works – which, as explained in a previous footnote means that the decay rate is likely to be underestimated. One might also want to cautious about extrapolating to the behaviour of current and future creative output from data of such elderly vintage.[23]

Liebowitz and Margolis (2005) argues that overall decay rates may be misleading and presents evidence that books that are popular upon release as measured by being bestsellers survive well (for example the table on p. 455 indicates that of the 91 bestsellers in their sample from the 1920s 54% are still in print 58 years later compared

[23]The issue of technological change is clearly an important one here: one might argue that with improvements in technology, both in production but also in distribution and discovery, the decay rate will fall in future. For example, it has been argued recently that technologies such as the Internet have made it easier to discover and access more obscure works leading to the growing importance of the 'long-tail' and a flattening of the distribution of sales (traditionally sales for most types of copyrightable goods have been dominated by a top 10-20% of works. The 'long-tail' then refers to the tail of this sales distribution). Here we do not explicitly consider the impact of technological change but we note that an earlier section dealt specifically with this issue.

to only 33% of non-bestsellers. However it is not clear how one should interpret this sort of evidence.

Simple 'in-print' status of a book only places a lower-bound on sales (furthermore a lower bound that is dropping with advances in technology) and does not allow us to compare the sales of a book today compared to when it was first released. More fundamentally, much heterogeneity is eliminated by the aggregation of copyrights into portfolios by the investors in creative work such as publishers, music labels and movie studios. In this case returns will tend to the average. Furthermore, were such aggregation not to occur it would require a substantial increase in the discount rate to take account of the increased uncertainty due to the reduction in diversification of the portfolio.[24]

5.7.4 Deadweight-Loss, Welfare Under Copyright and $\theta(n)$

Our preference would be to estimate all of these values directly from empirical data. However, this is daunting task given currently available datasets as it requires us to determine: the full demand system for copyright goods *and* the supply function for creative work. Because this task presents such insurmountable difficulties given present data availability we instead take a 'reduced-form' approach where we supply particular functional forms for the various quantities of interest (the average deadweight loss, marginal welfare etc). Where possible we calibrate these using existing data and we also perform robustness checks to ensure these results are reasonably robust. We begin by making the following assumptions:

1. The elasticity of production with respect to revenue, $s(n)$, is constant, equal to s.

2. The ratio of deadweight-loss to welfare under copyright on any given work is constant. This constant will be termed α.

3. The ratio of marginal welfare, $y(j)$, to marginal sales is constant. That is welfare follows the same trend as sales. This constant will be termed β.

[24] In these circumstances the issue of serial correlation would also become important. With high serial correlation, that is older successful works are those that were successful when young (and vice versa), the revenue when one extends term goes primarily to the owners works which have already generated substantial revenue (think here of a group like The Beatles). If one makes the standard assumption of diminishing marginal returns to creative output with respect to revenue, then serial correlation implies a very low elasticity of supply with respect to revenue – the revenue from extending term goes to those whose incomes are already high and therefore from whom little extra 'creation' can be expected when their incomes increase.

Assumption 1: little if anything is known about how the elasticity of supply with respect to revenue varies with the number of works produced. Furthermore we are already allow changes in welfare per work so it seems reasonable to take elasticity as constant.

Assumption 2: this assumption is questionable as one might expect that deadweight losses relative to welfare (under copyright) increase as the welfare (and revenue) from a work decline.[25] If this were so then this assumption would be incorrect and would result in an underestimate of the costs of copyright – and hence an overestimate of optimal copyright term. Nevertheless, we shall make this assumption for two reasons. First, it is difficult to derive estimates of this ratio from existing data. Second, as we shall see below, even with it (and the associated upward bias) we find that optimal term is well below the copyright terms found in the real world.

Assumption 3: this requires that the ratio welfare (under copyright) arising from a new work to the sales of that work does not vary over works. Again this is almost certainly not an accurate description of reality but as a first order approximation we believe it is not that bad. Furthermore, this assumption is crucial for our empirical strategy since it is relatively easy to obtain sales data compared to welfare data (which requires information on large segments of the demand curve).

Now, to proceed with the empirics, first let us switch to total welfare for notational convenience and define Y(j) to be total welfare under copyright from j new works so that $y(j) = Y'(j)$. Also define $Q(j)$ as total sales and $q(j) = Q'(j)$ as marginal sales (i.e. sales from the nth work).

What form does $Q(j)$ take? We shall assume it takes a 'power-law' form:

$$Q(j) = Aj^{\gamma}$$

This functional form appears to represent a reasonably good fit for sales of cultural goods and is frequently used in the literature.[26] We then have:

Lemma 5.7.2. *$\theta(n)$ has the following simple form:*

$$\theta(n) = \frac{\alpha}{s\gamma}$$

Proof. Recall that $\theta(n) = \frac{\overline{z}(n)}{sy(n)}$. Now $y(n) = \beta q(n)$, $z(n) = \alpha y(n)$ so average dead-

[25] For example, this would be the case if there was some fixed lower bound to transaction costs.

[26] See e.g. Goolsbee and Chevalier (2002); Ghose, Smith, and Telang (2004); Deschatres and Sornette (2004)

weight loss, $\overline{z}(n)$ equals $\alpha\frac{\beta Q(n)}{n}$. Hence:

$$\theta(n) = \frac{\alpha}{s}\frac{\beta Q(n)}{n\beta q(n)} = \frac{\alpha}{s\gamma}$$

□

Thus, one very convenient aspect of using a 'power-law' form is that $\theta(n)$ is not a function of n – it is 'scale-free'. In this case calculations of optimal copyright term do not depend on, n, the production function for works but only on α, γ and s.

5.7.5 Optimal Copyright Term: Point Estimates

Combining estimates of the ratio of deadweight losses to welfare under copyright (α) and the rate of diminishing returns (γ) with those provided above for cultural decay (b) and the discount factor (d) we will obtain point estimates for optimal copyright term.

s

There is an almost total lack of data which would allow us to estimate the elasticity of supply with respect to revenue. Landes and Posner (2003) who point out that there is no discernible impact on output of work from the US 1976 extension of term. Hui and Png (2002) find a similar result when looking at movies and the CTEA in the US though more recent work with a cross-country dataset, Png and hong Wang (2007), does find an impact. Given this uncertainty and lack of information the best we can do is to posit what we feel is a plausible range for s of $[0.5, 1.5]$ with an average value of 1.0.

γ

Ghose, Smith, and Telang (2004) list a whole range of estimates for $\gamma - 1$ (all derived from Amazon) ranging from -0.834 to -0.952 with the best estimate being -0.871. These imply γ in the range 0.048 to 0.166 with best estimate at 0.129. We shall proceed using this estimate of 0.129.

α

Estimating α is harder because of the paucity of data which would permit estimation of off-equilibrium points on the demand curve. However the available evidence though scanty suggests that the ratio could be quite large. For example, Rob and Waldfogel

(2004) investigate file-sharing among college students and estimate an implicit value for deadweight-loss of around 36% of total sales. Converting this to welfare ratio requires some assumption about the ratio of welfare (under copyright) to sales. A linear demand structure (with zero marginal costs) would give a deadweight loss to sales ratio of 50% and deadweight loss to welfare under copyright ratio of a third. Increasing marginal costs would reduce the ratio to sales but keep the ratio to welfare constant at a third. Being more conservative, assuming producer surplus were around 50% of sales and consumer surplus to be two to five times would give a value for α of between 0.24 and 0.12. Other papers, such as Le Guel and Rochelandet (2005); Ghose, Smith, and Telang (2004), while not providing sufficient data to estimate deadweight loss, do suggest it is reasonably substantial. Thus, we feel a plausible, and reasonably conservative, range for α would be from $[0.05, 0.2]$, that is deadweight loss per work is, on average, from a twentieth to a fifth of welfare derived from a work under copyright. When required to use a single value we will use the halfway point of this range 0.12.

5.7.6 A Point Estimate for Optimal Copyright Term

With $\alpha = 0.12, s = 1.0, \gamma = 0.129$ then $\theta \approx 0.93$. With our defaults of a discount rate of 6% and cultural decay of 5% this implies an **optimal copyright term of around 15 years**.

5.7.7 Robustness Checks

Given the uncertainty over the values of some of the variables it is important to derive optimal copyright term under a variety of scenarios to check the robustness of these results. Table 5.1 presents optimal term under a range of possible parameter values including those at the extreme of the ranges suggested above.

With variables at the very lower end of the spectrum (the first row) optimal term comes out at 52 years which is substantially shorter than authorial copyright term in almost all jurisdictions and roughly equal to the 50 years frequently afforded to neighbouring rights (such as those in recordings). However as we move to scenarios with higher levels for the exogenous variables optimal term drops sharply. For example, with cultural decay at 3%, the discount rate at 5% and the ratio of deadweight loss to welfare under copyright at 7% we already have an optimum term of just over 30 years. At the very highest end of the spectrum presented here, with deadweight losses at 20% of welfare under copyright (recall that a linear demand curve corresponds to a 33% ratio) and cultural decay and the discount rate both at 8% optimal term is around four and a half years.

Cultural Decay Rate (%)	Discount Rate (%)	α	Optimal Term
2	4	0.05	51.97
3.5	5	0.07	30.63
5	6	0.1	18.06
6.5	7	0.15	9.25
8	8	0.2	4.53

Table 5.1: Optimal Term Under Various Scenarios. α is the ratio of deadweight loss to welfare under copyright and s (the elasticity of supply) is set to 1 and γ (sales curve exponent) to 0.129.

We can also plot a probability density function under the assumed variable ranges. This has the advantage that incorporates the interrelations of the various variables – by contrast, Table 5.1, by nature of its form, implicitly gives the inaccurate impression that each of the outcomes listed is equally likely. We present the distribution function in Figure 5.1. As this shows, the mode of the distribution is around 20 years and the median is just under 15 years. From the underlying cumulative distribution function we can calculate percentiles and find the 95th percentile at just under 31 years, the 99th percentile at 39 years and the 99.9th percentile at just over 47 years. This would suggest, that at least under the parameters ranges used here, one can be extremely confident that copyright term should be 50 years or less – and it is highly like that term is under 30 years (95th percentile).

An Inverse Approach

An alternative approach to estimating underlying parameters and using that to find the optimal term is to look at the inverse problem of calculating the 'break-even' value for a particular variable for a given copyright term. The 'break-even' value is the level of that variable for which that term is optimal. Here we will focus on α, the ratio of deadweight loss to welfare under copyright – so if the actual value α is higher than this break-even level then term is too long and if actual α is below it then term is too short. This provides a useful robustness: derive the break-even α corresponding to the copyright term currently in existence and then compare this value to whatever is a plausible range for α. If the value is outside this range one can be reasonably certain that current copyright term is too long.

Given our assumption on the form of the discount factor and the rate of cultural decay theta takes the following form:

$$\alpha^{-1}(S) = \frac{d^S(1-bd)}{\frac{1-b}{s\gamma} - d^S b^S(1-bd)}$$

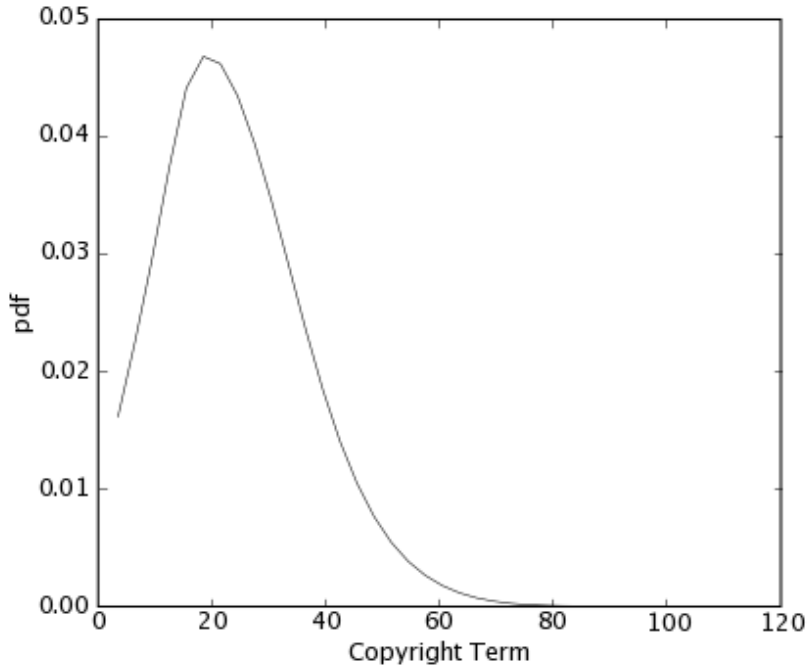

Figure 5.1: Probability distribution of optimal term given the parameter range set out above (with the exception that γ takes a single value of 0.129).

Figure 5.2 provides a plot of this inverse, 'break-even', function. Under the Berne convention minimal terms of protection for most types of work is life plus 50 years (and many countries including the US and all of those in the EU now provide for life plus 70). This in turn will correspond to a copyright length of somewhere between 70 and 120 years (assuming the work is created between the ages of 20 and 70). Let us take a low value in this range, say 80 years. We summarize the 'break-even' α corresponding to term of this length in Table 5.2 focusing on a set of very conservative parameter values. As can be seen there, even with a cultural decay rate of 2%, a discount rate of 4% and elasticity at its uppermost value the break-even α is 2.5% – so for any α above that term is too long. With a slightly higher decay and discount rate (3% and 5% respectively) break-even α falls to 1.3%. Thus, even with low values for the discount and cultural decay rate the level of α required for current copyright terms to be optimal seem too low to be plausible.

5.8 Conclusion

In this paper we have developed a simple framework for analysing copyright grounded in the existing literature. Using it, we obtained two sets of separate, but complementary,

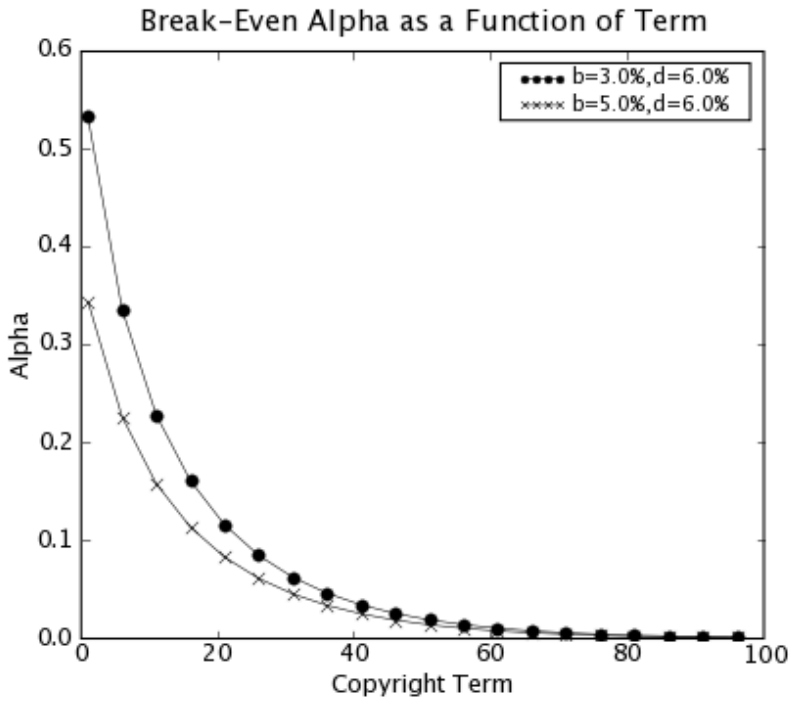

Figure 5.2: Break-even alpha as a function of copyright term. b is the cultural decay factor and d the discount factor.

Cultural Decay Rate (%)	**Discount Rate (%)**	**Elasticity**	**Break-even α (%)**
2	4	1.5	2.5
2	4	1.0	1.6
3	5	1.5	1.0
3	5	1.0	0.7

Table 5.2: Break-even α (per work ratio of deadweight loss to welfare under copyright). γ is set to 0.129.

results. In the first section, we investigated the effect of changing production costs on the optimal level of copyright as well as how the level of optimal protection varies over time. We demonstrated in substantial generality that (a) optimal protection is likely to fall with a decline in the costs of production and distribution of 'originals' (b) in contrast to the presumption of some existing policy making, technological change which decreases costs, because it effects both 'originals' and 'copies' may imply a decrease as well an increase in optimal copyright (c) under the reasonable assumption that the stock of 'effective' work is non-decreasing the level of optimal copyright falls over time.

In the second section we turned our attention to one specific aspect of optimal copyright, namely the term of protection. In Theorem 5.7.1 we used our model to derive a single equation that defined optimal term as a function of key exogenous variables. Using the estimates for these variables derived from the available empirical data we obtained an estimate for optimal copyright term of approximately 15 years. To our knowledge this is one of the first estimates of optimal copyright term which is properly grounded, both theoretically and empirically, to appear in the literature.

All our results have significant implications for policy. In recent times technological change has substantially reduced the costs of production and distribution of most copyrightable goods. Much of the existing policy discussion has focused, almost exclusively, on reductions in the costs of 'unauthorised' ('pirate') copies and has tended to assume that this necessitates an increase in the level of protection. However, as we pointed out, the costs of 'originals' have also fallen dramatically, and this change is likely to require a *reduction* in the strength of protection. Looking more generally at the case of technological change which reduces the production costs of both 'originals' and 'copies', the implications for copyright policy were ambiguous – not surprising given the two contrary effects at play – and we highlighted the key terms in need of empirical estimation if an unambiguous answer were to be obtained.

Moving on we came to the question of optimal copyright term – probably the most important aspect of the overall 'level' of copyright. Our estimate of optimal term (15 years) is far below the length copyright in almost all jurisdictions and we confirmed this general fact – that current copyrights are likely too long – using several robustness checks. This implies that there is a significant role for policymakers to improve social welfare by reducing copyright term as well as indicating that existing terms should not be extended. Such a result is particularly importance given the degree of recent debate on this precise topic.

Finally, there remains plentiful scope to extend and build upon the work here. In particular, there is room for further empirical work on all aspects of these results. For example, it would be valuable to calibrate the production costs model to investigate

what changes in the level of copyright would be implied by the recent reductions in the cost of production and distribution. Similar work could be done in relation to changes of copyright over time where one would need to collect data on the level of production and the form of the welfare function.

Regarding the derivation of optimal term, the main challenge would be to improve the estimates for the key parameters, especially that of the ratio of deadweight loss to welfare under copyright. As discussed above, the perfect approach would involve estimating the demand-system for the copyrightable goods under consideration. This is a non-trivial task but one of great value – and not simply in relation to the problems considered here.

Bibliography

George A. Akerlof, Kenneth J. Arrow, Timothy Bresnahan, James M. Buchanan, Ronald Coase, Linda R. Cohen, Milton Friedman, Jerry R. Green, Robert W. Hahn, Thomas W. Hazlett, C. Scott Hemphill, Robert E. Litan, Roger G. Noll, Richard L. Schmalensee, Steven Shavell, Hal R. Varian, and Richard J. Zeckhauser. The Copyright Term Extension Act of 1998: An Economic Analysis, 5 2002.

David Blackburn. Online Piracy and Recorded Music Sales, 12 2004. Job Market Paper (Harvard PhD Programme).

Michele Boldrin and David Levine. IP and market size. Levine's Working Paper Archive 618897000000000836, UCLA Department of Economics, July 2005.

Tim Brooks. Survey of Reissues of US Recordings, 2005. Copublished by the Council on Library and Information Resources and the Library of Congress.

CIPIL. Review of the Economic Evidence Relating to an Extension of the Term of Copyright in Sound Recordings, 12 2006. Prepared for the Gowers Review on Intellectual Property.

William Cornish and David Llewelyn. *Patents, Copyrights, Trade Marks and Allied Rights*. Sweet and Maxwell, London, 5th edition, 2003.

F. Deschatres and D. Sornette. The Dynamics of Book Sales: Endogenous versus Exogenous Shocks in Complex Networks, December 2004.

A. Ghose, M.D. Smith, and R. Telang. Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact, 2004.

Austan Goolsbee and Judith Chevalier. Measuring Prices and Price Competition Online: Amazon and Barnes and Noble, July 2002. NBER Working Papers.

Kai-Lung Hui and I. P. L. Png. On the Supply of Creative Work: Evidence from the Movies. *American Economic Review*, 92(2):217–220, May 2002.

William Johnson. The Economics of Copying. *Journal of Political Economy*, 93(1): 158–174, 1985.

William Landes and Richard Posner. An Economic Analysis of Copyright Law. *Journal of Legal Studies*, 18(2):325–363, 1989.

William Landes and Richard Posner. Indefinitely Renewable Copyright. *University of Chicago Law Review*, 70(471), 2003.

Fabrice Le Guel and Fabrice Rochelandet. P2P Music-Sharing Networks: Why Legal Fight Against Copiers May be Inefficient?, Oct 2005.

Stan Liebowitz. Copying and Indirect Appropriability: Photocopying of Journals. *Journal of Political Economy*, 93(5):945–957, 1985.

Stan Liebowitz and Stephen Margolis. Seventeen Famous Economists Weigh in on Copyright: The Role of Theory, Empirics, & Network Effects. *Harvard Journal of Law and Technology*, 18(2), 2005.

William Nordhaus. *Invention, Growth and Welfare: A Theoretical Treatment of Technological Change*. M.I.T. Press, 1969.

Ian Novos and Michael Waldman. The Effects of Increased Copyright Protection: An Analytic Approach. *Journal of Political Economy*, 92(2):236–246, 1984.

Felix Oberholzer and Koleman Strumpf. The Effect of File Sharing on Record Sales: An Empirical Analysis. *Journal of Political Economy*, 115(1):1–42, 2007.

Martin Peitz and Patrick Waelbroeck. Why the music industry may gain from free downloading – The role of sampling. *International Journal of Industrial Organization*, 24(5):907–913, September 2006.

Ivan Png. Copyright: A Plea for Empirical Research. *Review of Economic Research on Copyright Issues*, 3(2):3–13, 2006.

Ivan Png and Qiu hong Wang. Copyright Duration and the Supply of Creative Work. Levine's Working Paper Archive 321307000000000478, UCLA Department of Economics, January 2007.

E. Rappaport. Copyright Term Extension: Estimating the Economic Values. Technical report, Congressional Research Service, May 1998.

Rafael Rob and Joel Waldfogel. Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students, 2004.

Nancy L. Stokey, Jr. Lucas Robert E., and Edward C. Prescott. *Recursive Methods in Economic Dynamics*. Harvard University Press, October 1989. ISBN 0674750969.

Richard Watt. *Copyright and Economic Theory: Friends or Foes?* Edward Elgar, Cheltenham, UK, 2000.

Chapter 6

The Control of Porting in Two-Sided Markets

6.1 Introduction

Several recent cases, which we discuss in more detail below, have focused economists' attention on the motivations and effects of the behaviour of a dominant firm in two-sided markets. We believe that much of this activity can usefully be interpreted in terms of efforts to control (and prevent) 'porting' – where porting denotes the conversion of a 'software' or 'service' associated with one platform to run on another platform. Building on the existing literature on two-sided markets, we develop a formal model of 'porting' and, focusing on the case where a dominant platform exists. We use this model to investigate the impact on equilibrium and the consequences for welfare of the ability to control porting. Specifically, we show that the welfare costs associated with the 'control of porting' may be more significantly more substantial than those arising from pricing alone.

For example, much of the 1998 case of *US vs. Microsoft* as well as more recent antitrust disputes in Europe over Microsoft's media player can be seen as related to efforts to control porting. In the 1998 case there was the alleged tying of Internet Explorer browser as well as efforts to undermine compatibility with other systems, for example, by subtly changing the Windows version of the Java Virtual Machine.[1] Similarly, the media player dispute concerned the bundling of Microsoft's own Media Player 'for free' with the operating system. In both cases there has been considerable debate[2] over the motivations for, and consequences of, Microsoft's behaviour, especially as to whether these sorts of activities could be described as 'tying'.[3] To our mind much of this behaviour is best seen in light of efforts to control porting and thereby preserve the market power associated with the 'Applications Barrier to Entry' (as the indirect network effects were termed in that anti-trust action). Unlike with traditional tying, Microsoft's actions, though obviously directly *affecting* competing applications (Netscape's Browser, Real Networks Audioplayer etc), were not *directed* at them. Rather, they were motivated by the fear that losing control of key Application Programming Interfaces (APIs) and user services would make it easier for end-user applications and services to move (port) between operating system platforms, which would, in turn, make it easier for consumers to switch between different platforms and thereby reduce Microsoft's market power.[4]

[1] See Judge Jackson in *Findings of Fact in the case of United States vs. Microsoft*, (Jackson, 1999).

[2] See, for example, Hall and Hall (2000); Davis and Murphy (2000); Fisher (2000); Bresnahan (2001); Liebowitz and Margolis (1999); Klein (2001); Gilbert and Katz (2001).

[3] See the works previously cited and, specifically on the tying issues, Whinston (1990); Bernheim and Whinston (1998) and the survey in Whinston (2001).

[4] This also explains why Microsoft only 'integrates/ties' certain applications and is happy for most software to be produced by third-party vendors. The need to tie only arises when that application or service will itself be the site of significant third-party development. This is clearly the case with web-

Another example is provided by the 2000 case of eBay vs. Bidder's Edge.[5] Here, eBay, an online auction site, successfully sued Bidder's Edge, a firm which collected together prices from different auction site for consumers to compare, for cyber-trespass, ostensibly on the grounds that Bidder's Edge spidering activities caused excessive load on their servers. However, as various commentators pointed out the ability to exclude a firm such as Bidder's Edge could also have serious anti-competitive effects[6]. EBay is a classic example of a platform in a two-sided market with sellers taking the role of 'software' or 'service' and buyers that of consumers. If a third-party were easily able to transfer (port) sellers from one auction platform to another then eBay's market power would be greatly diminished. A firm such as Bidder's Edge would greatly facilitate such 'porting' by ensuring that a given seller (and their associated 'reputation') would be visible to consumers no matter what auction platform they were on. By preventing Bidder's Edge (and any other similar firm) from being able to extract data from the eBay site without permission eBay obtained very substantial control of porting from its platform.

A final example comes from the ongoing debate in Europe around interoperability of TPMs/DRMs (Technological Protection Measures/Digital Rights Management) systems, particularly in relation to the dominant position of Apple's iPod and iTunes products both of which use Apple's proprietary 'FairPlay' DRM. Here the platform is the digital music player and the 'software' is the music. Apple operates on both sides of the market with the iPod or iTunes software on the platform side and the iTunes Music Store (ITMS) on the 'software' (music) side. If DRM were interoperable then one could play a song from any given digital music store on any given digital music player. However with no interoperability if someone buys all their songs from the iTunes Music Store (currently with 70-80% of the digital downloads market) then they can only play them on an iPod (and if they change music player they may lose all their purchased music). Thus proprietary DRM makes it substantially harder for consumers to switch platforms (i.e. digital music players). By maintaining a closed, proprietary, DRM system and integrating backwards into the 'software' (music) market (analogously to the previous Microsoft examples) Apple are able to effectively control porting and thereby increase their market power in the platform (music player) market.[7]

---

browsers, as Bill Gates presciently saw in his 'Internet Tidal Wave' memorandum: "A new competitor 'born' on the Internet is Netscape [Netscape was launched 15th Dec 1994]. Their browser is dominant with 70% usage share, allowing them to determine what network extensions will catch on. They are pursuing a multi-platform strategy where they *move the key API into the client to commoditize the underlying operating system* ..." (emphasis added).

[5]EBAY, Inc vs. BIDDER'S EDGE Inc, http://pub.bna.com/lw/21200.htm.

[6]See, for example, the amicus curiae brief filed by a collection of 28 law professors available online at http://www.gseis.ucla.edu/iclp/ebay-ml.

[7]It is important to note for this analysis that it is well-known that Apple make their profits on the

The paper builds upon several strands in the existing literature. First, there is existing work on 'converters' in network markets (converters being devices that allow a user on one network to gain access to a separate network). For example, Farrell and Saloner (1992) examine the provision and purchase of imperfect converters in a network effects model, as well as the incentive for a dominant firm to make conversion costly.[8] As porting can be seen as the analogous activity in a two-sided market with 'indirect network effects' to converters in the original 'one-sided' models our work can be seen as extending this existing work to the more complex two-sided case.

The second strand is the literature on indirect network effects and two-sided markets. Early work by Church and Gandal (1992) (extended by Church, Gandal, and Krause (2003)) analyzed the case where consumers cared about the variety of complementary goods available for a particular platform or network. They showed that with fixed costs in production this led to 'indirect network effects', that is a positive relation between the utility of a consumer from a given platform and the number of other consumers joining that platform. This work has recently been extended and generalized under the heading of two-sided markets, see, for example, Armstrong (2005); Rochet and Tirole (2003, 2005). The focus of much of this literature has been on the charging decisions of the platform owner – in particular, the form of fees and what determines the level of fees, and subsidies, on the two sides (the 'software' side and the consumer side). By contrast, in this paper we are interested in something rather different: what happens if one platform owner can influence the availability of 'software' on the other platform by controlling porting (that is the ability of 'software' to multi-home).

Seen in this light, the closest work to ours in the existing two-sided literature are the papers of Armstrong and Wright (2005) and Choi (2006). Armstrong and Wright (2005) provides a general examination of two-sided markets with multi-homing. In particular, they consider the use of exclusive contracts by a platform owner as means to force single-homing on the seller side. However, due to the complexity of the analysis in the full two-sided the case the authors fall back to analyzing the case of pure network effects.[9] Our model differs from this in several ways. First, rather than exclusive contracts we have a general 'porting cost' variable which influences the ability of 'software' produced for one platform to move to the other. Second, we allow for ex-ante asymmetry in platform's market share and general forms for both heterogeneity and indirect network effects. However, like Armstrong and Wright (2005), the fully general case is too

hardware (the iPod) and make very little from the iTunes Music Store.

[8]See also Choi (1997) for another converter model, albeit a dynamic one related to the transition from an old to a new technology.

[9]The focus is on the case of symmetric platforms which may be problematic when analyzing tying, as the authors state (p. 22): 'Given the underlying symmetry of firms in our model, it is not obvious that exclusive contracts are advantageous to the platforms in equilibrium.'

complex for ready analysis and so the price we pay to keep the model tractable is some degree of restriction on platform's pricing decisions as well as confining ourselves to the case where a single proprietary (and dominant) platform faces a competitive one.

Choi (2006) presents a rather different model, which is primarily animated by the Media Player case, focusing on the combination of tying with multi-homing on the buyer (consumer) side. Here, tying is about the ability for a firm with a monopoly in some underlying market to use tying to monopolize a related two-sided platform (for example Microsoft using its operating system monopoly to control the media players). With multi-homing on the buyer side Choi finds that the welfare effects of tying are ambiguous with tying in some cases being welfare improving. Our concerns are rather different. First, we have 'porting' (multi-homing) on the 'software' (seller) side, not the consumer (buyer) side, of the market. Second, and more importantly, the 'tying' in our model is between the platform and its associated 'software', not between some outside product and the platform.

Finally, our paper obviously has commonalities with the literature on tying and vertical foreclosure (see e.g. Whinston (1990); Bernheim and Whinston (1998)). Due to the prominence of the tying issue in the Microsoft case there has been a flurry of papers on tying models. Perhaps the closest, at least in spirit, to the model presented here is that of Gilbert and Riordan (2007) who investigate what they term 'technological' tying by a monopolist. Increasing porting cost in our model could be seen as analogous to the 'technological' tying in their model (whereby the quality of a complementor can be reduced by the monopolist). That said, technological tying is similar to traditional tying in that it is motivated by a desire to sell the complementary good (or the bundle), whereas for the case of porting examined here that is so: the monopolist simply wishes to inhibit complementors from porting to another platform in order to reduce competition with *its own platform*. Integration, if it happens at all, may occur not because it is profitable in itself – it may even be loss-making – but only because it reduces the degree of platform competition.

6.2 The Model

The basic framework is that used in the two-sided markets literature (see e.g. Armstrong (2005)). There are two platforms/networks: $X = A, B$ and a mass of consumers (buyers) modelled by the interval $[0, 1]$ with the index, $t \in [0, 1]$, used to label them. The measure of consumers on platform X is denoted by n_X. Each platform has an associated set of 'software/services' ('sellers') and the amount of 'software' available on platform X is s_X.

Consumers derive utility from using software and must purchase access to a platform to be able to use the associated software.[10] Consumers are heterogeneous in their preferences for a given platform.[11] If a consumer has already purchased 'software' from one platform she gains no extra utility from purchasing from a second platform so a consumer will purchase from at most one platform (there is no multi-homing on the buyer side). We also make the standard assumption that all consumers join one or other platform.

Formally, consumers have the following utility function:

$$u_X(t, p_X, s_X, p_X^s) = \phi - p_X - h_X(t) + u_X^s(s_X, p_X^s)$$

Where

- ϕ is a positive constant introduced so that reservation utility can be normalized to 0 (alternatively one could remove ϕ from utility function and set reservation utility to $-\phi$)
- p_X is the price of hardware on platform X
- $h_X(t)$ models consumer heterogeneity. It is assumed that heterogeneity is symmetric across platforms that is, $h_B(1-t) = h_A(t)$. This allows one to write $h_A(t) = h(t) = h_B(1-t)$. We shall assume the standard 'orderability' of consumers by heterogeneity, i.e. $h'(t) > 0$. Thus we have a standard linear city model with platform A at 0 and platform B at 1 and consumers preferring, all other things being equal, a closer platform.
- u_X^s is utility from software purchases with s_X the amount of software available on platform X and p_X^s the price (or vector of prices) of software. This is discussed further below.

Platform A is controlled by a single firm, the monopolist (M). Platform B is provided competitively. Platform fixed costs are assumed to be sunk and therefore may be taken without loss of generality to be zero. Marginal costs of access per consumer, c, are constant and the same for each platform. Since platform B is perfectly competitive the access price equals marginal cost: $p_B = c$. Since the marginal cost is common across the two platforms we may, without loss of generality, set $c = 0$.

[10] There are no 'direct' network effects, that is consumers' utility from a given platform is affected directly only by the platform itself and the amount of software available on it and not by the number of other consumers using that platform. It would not be difficult to incorporate direct effects into our model but as that is not the focus of our analysis here we have chosen to omit them for the sake of simplicity.

[11] This could be taken as encapsulating general differences in the type of software available on the two platforms.

6.2.1 Software Production and Porting

The software that is produced may be created by two methods. Either it can be created directly for platform X at fixed cost f_X^d or it can be ported from the other platform at fixed cost f_X^p (note that this only relates to the fixed cost, the marginal cost is the same whether the software is ported or created directly). In what follows the main focus will be on the cost of porting from the monopolist's platform (A) and so we will drop the subscript and define $f^p = f_A^p$.

In our model we will suppose that a monopolist may increase the cost of porting from its platform to a competitor's – though at the cost of some expenditure on its own part.[12] Formally, if e is expenditure then $f^p = f^p(e)$. It will be convenient in what follows to have the porting cost, f^p, being the choice variable rather than expenditure, e. This simply involves using the inverse function (the expenditure to prevent porting), $e = e(f^p)$. Efforts to prevent porting display diminishing returns so $e'(f^p) > 0, e''(f^p) > 0$.

Thus the fixed cost of software production on a platform, f_X, will be either: f_X^d if all software is produced directly (none is ported); a mixture of f_X^d and f_X^p if some software is ported and some produced directly; or f_X^p if all software is ported.

6.2.2 Sequence of Actions

1. The monopolist, M, chooses values for control variables: p_A, f^p.

2. Software producers for a given platform form expectations of platform size. Based on these expectations, producers decide whether to engage in direct software production. Then, given this amount of direct production, (other) producers[13] decide whether to engage in porting of this existing, directly produced, software.

3. Taking the resulting level of software provision and prices as given consumers solve their utility maximization problem and decide from which platform to purchase.

4. M's profits, $\Pi = p_A \cdot n_A(p_A, f^p) - e(f^p)$, are determined.

[12] For motivation the reader is directed to some of the examples set out in the introduction with perhaps the most relevant one here being the behaviour of Microsoft. Microsoft has incurred significant expenditure on several products, e.g. its Java Virtual Machine, Internet Explorer, the .NET framework, and Windows Media Player, where it would appear that a substantial motivation for the products development was the desire to make it harder to port software and services from its own system to competitors (in each case the product increased Microsoft's control of key APIs and formats).

[13] Here it will not matter whether the firms that engage both in direct production and porting are the same or different since direct production confers no special ability in porting and, as with free entry, in equilibrium all producers earn zero profits.

Remark 6.2.1. In equilibrium the resulting platform sizes must be consistent with rational expectations. That is: actual and expected platform sizes are equal and actual and expected software levels are equal. In this case the order in which software firms and consumers move does not affect the outcome of the model. Thus we could as easily have software firms taking their decisions after consumers or even simultaneously.

6.3 Solving the Model

We take a general approach in which we assume only that software production on platform X involves (a) some form of fixed costs (f_X) (b) that the amount and price of software on platform X may be expressed solely in terms of these fixed costs, f_X and the number of consumers on the platform, n_X. Taken together these mean that the consumer software utility function has a reduced form of the following kind:

$$\begin{aligned} u_X^s(s_X, p_X^s) &= u_X^s(s_X(f_X, n_X), p_X^s(f_X, n_X)) && (6.1) \\ &\equiv \nu_X(f_X, n_X) \text{ with } \nu_{f_X} < 0, \nu_{n_X} > 0 && (6.2) \end{aligned}$$

We shall term ν_X the 'indirect network effects' function on platform X.[14] By proceeding in this manner the results are kept as general as possible. Furthermore, the two basic models of imperfect competition with fixed costs (monopolistic competition and product differentiation) can both be shown to give rise to this reduced form (Appendix 6.B provides an explicit derivation for the case of a standard circular city model of product differentiation).

As presented we now have a standard two-sided model with utility functions:

$$u_X(t, p_X, f_X, n_X) = \phi - p_X - h_X(t) + \nu(f_X, n_X)$$

We can solve this in the usual manner to obtain platform sizes as a function of the monopolist's choice variables: $n_A = n_A(p_A, f^p)$.[15] The monopolist then solves:

$$\max_{p_A, f^p} p_A n_A(p_A, f^p) - e(f^p)$$

[14]Note that we implicitly assume some symmetry across platforms in that the function ν is the same for the two platforms.

[15]n_A will also depend on other variables such as the direct cost of software production but these are exogenous variables not under the control of any player.

6.3.1 Solving for the Subgame Equilibrium

We solve first for the equilibrium platform size in the consumer/software subgame (stage 2 onwards, that is after M has set prices and porting cost). We proceed by the usual method based on finding the marginal consumer indifferent between the two platforms.

First, recall that we have assumed that consumers gain no extra utility by purchasing from more than one platform. Thus, we may assume that consumers purchase at most one platform. We also assumed that all consumers do purchase from one or other platform. Thus we have $n_B = 1 - n_A$ and we need only consider n_A in what follows. For notational convenience suppress auxiliary variables in the consumer utility functions and write $u_X(t, p_X, f_X, n_X) = u_X(t, n_X)$.

Define: the *conditional utility advantage* of platform A over platform B for consumer t when platform size is n_A:

$$\hat{A}(t, n_A) = u_A(t, n_A) - u_B(t, 1 - n_A)$$

and the *utility advantage (function)*, which gives the utility advantage of platform A over B if t is the marginal consumer (so $t = n_A$):

$$A(t) = \hat{A}(t, t)$$

Using the expression for the utility function we have that:

$$A(t) = -p_A - h_A(t) + h_B(t) + \nu(f_A, t) - \nu(f_B, 1 - t)$$

Lemma 6.3.1. *The equilibria of the subgame from stage 2 onwards (after M sets price and porting costs) are given by $E = E_0 \cup E_{-0}$ where E_0 is the set of interior equilibrium, $E_0 = \{t : A(t) = 0\}$, and E_{-0} is the set of extremal or 'standardization' equilibrium in which all consumers join one or other platform, $E_{-0} = \{0 : A(0) < 0\} \cup \{1 : A(1) > 0\}$. An equilibrium $t_e \in E_0$ is stable if $A'(t_e) < 0$. All $t_e \in E_{-0}$ are stable.*

Proof. See appendix. □

Note that the advantage function implicitly depends on all of our exogenous and choice variables: $A(t) = A(t, p_A, f_A, f_B)$ and therefore so does the set of equilibria $E = E(p_A, f_A, f_B)$. We make the following assumption about the existence of an equilibrium to this subgame:

Assumption: The exogenous variables, in this case the functional forms for the heterogeneity and indirect network effects, are such that, when $p_A = 0$ and the porting

cost at its initial value (that is without any intervention by M), there would exist an asymmetric stable interior equilibrium where platform A is larger than B.

Justification: without a stable equilibrium of the subgame the overall game will clearly have no equilibrium. Thus we must have the existence of at least one stable equilibrium of the subgame.[16] We require the existence of an *interior* stable equilibrium to the subgame for two reasons. First, in most real-world scenarios, even those that involve a very dominant platform, we rarely see a platform with 100% market share. Second, at an extremal equilibria the monopolist actions no longer have a marginal impact (for example, the monopolist may increase or decrease prices without any impact on demand). This renders such equilibria both less interesting and more cumbersome to analyze. Finally, with regard to the asymmetry: in most real world situations one platform is larger than the other. Furthermore, in any situation with antitrust considerations this will necessarily be the case.

6.3.2 Porting

In this section we shall determine the amount of software produced for each platform of the various possible types (produced directly, ported or produced by a mixture of those methods). In doing so, we will also have determined the 'actual' fixed cost of software production for each platform f_A, f_B in terms of the fixed cost of directly producing software for that platform and the (common) porting cost (f_X^d, f^p). To simplify the statement of results it will be useful to make a technical assumption to exclude one particular measure zero configuration of (expected) platform sizes and direct software production costs:

Technical Assumption: $f_A^d n_B \neq n_A f_B^d$.

Lemma 6.3.2 (Porting Lemma)**.** *In equilibrium only one platform has software produced directly for it. All the software on the other platform derives from porting. Let X denote the platform for which software is produced directly and denote the other by X'. Then $f_X = f_X^d$ and the amount of software on X' will be equal to the smaller of:*

1. *The amount of software on X (in the case where all software is ported)*
2. *The level of software production is determined by the porting cost, i.e. the level of software production is that which would be produced with $f_{X'} = f^p$.*

[16] This part need not be an assumption since under mild conditions, such as symmetry of the indirect network effects and heterogeneity function, one can show there exists at least one stable equilibrium to the subgame. However it is clearly not possible to ensure the existence of a stable *interior* equilibrium in general – consider the standard symmetric case with linear heterogeneity and network effects: he only interior equilibrium is at 0.5 and with 'strong' network effects this equilibrium must be unstable.

If the second case obtains, i.e. not all possible software is ported (so the level of porting cost matters), the porting constraint will be said to bind and we have $f_{X'} = f^p$.

Proof. See appendix. □

Remark 6.3.3. The result that, for any given platform, all software is either produced directly or ported may seem a little implausible. After all, in reality, we usually see software produced directly for all platforms. It is also usual for there to be substantial porting, with the same piece of software available on multiple platforms (multi-homing on the software side).[17] However, all that is necessary for the results in this paper is that the *marginal* piece of software on the platform competing with the monopolist is ported – in which case altering the costs of porting change the amount of software on that platform. Thus, while the model as given may seem to be overly restrictive in its implications the necessary result, that is that the porting constraint binds, will still hold in the more general case.

We now make one further assumption:
Assumption: In the case of asymmetry, it is the platform with larger (expected) size for which software is produced directly.

Justification: we have just shown that it will always be the case (in this model) that software on one platform has all software produced directly and one has all software ported. Since the amount of software on the 'porting' platform must always be less than or equal to that on the 'direct-production' platform it is natural to assume that it is the platform with larger (expected) size for which software is produced directly.[18]

Combining these assumptions with the results of the previous section we may set $f_A = f_A^d$ and $f_B = f^p$ (though we will need to check that the porting constraint does not bind).

6.3.3 Solving for Overall Equilibrium

Finally it is necessary to demonstrate the existence of an equilibrium in the overall game: that is a solution to the monopolist's profit maximization problem taking account of how the monopolist's choices impact on the actions that will be taken by other agents (consumers and software producers). This response of other agents to M's choices has already been derived in the form of the subgame 'network' equilibrium derived above. We note that these results may not be easy to grasp when presented as

[17]Extending the model to have direct production on both platforms and intermediate levels of of porting could most easily be done by allowing heterogeneity in both direct production and porting costs of 'software'.

[18]In fact if platforms displayed symmetry, i.e. direct production costs are equal and heterogeneity functions on the two platforms are the same, this is a result rather than an assumption.

generally as they are here. The following section examines a specific case graphically and the reader may find it profitable to peruse that example first before returning to the more abstract approach used here.

Lemma 6.3.4. *Having picked a stable interior equilibrium $t_e^0 \in E_0(p_A^0, f_A, f_B)$ we have associated to it a well-defined, continuous and differentiable 'equilibrium function', $t_e(p_A, f_A, f_B)$, defined in a neighbourhood of t_e^0. In particular, restricting to changes in p_A we have a demand function:*

$$q(p_A) = t_e(p_A) = A^{-1}(p_A)$$

Differentiating we have:

1. *Downward sloping demand schedule: $\frac{dq}{dp_A} = \frac{-1}{A'(t_e(p_A))t_e'(p_A)} < 0$*
2. *$\frac{dt_e}{df_A} < 0$*
3. *$\frac{dt_e}{df_B} > 0$*

Finally, though demand may be discontinuous at some point, there exists locally, that is within the region where demand is continuous, a unique profit maximizing price.

Proof. See appendix. □

Combined with the results of the previous section the monopolist's profit maximization problem becomes:

$$\max_{p_A, f^p} p_A \cdot t_e(p_A, f^p) - e(f^p)$$

We make one final additional technical assumption which allows us to rule out the possibility of discontinuities in M's profit function as a result of changes in porting cost:

Assumption: Pick such an asymmetric stable interior equilibrium t_e^0 and consider the associated equilibrium function $t_e(p_A, f^p)$. Then that function exists and is continuous for all values of f^p up to f_B^p (which is the maximal value that f^p would ever be set to by M).

Corollary 6.3.5. *There exists an equilibrium of the overall game, that is a price and porting cost and an associated equilibrium level of demand $t^e(p_A, f^p)$ which maximize the monopolist's profits.*

Proof. See appendix. □

6.3.4 Example I: Equilibrium and Demand

The situation we shall consider is one in which the two platforms are a priori equivalent, that is the fixed costs of software production on the two platforms are equal and heterogeneity is symmetric ($h_B(1-t) = h_A(t)$).[19] For the 'network effects' function we use the reduced form derived from a circular city model (see appendix), that is $\nu(f,t) = C - \sqrt{\frac{f}{t}}$. This form differs substantially from the classic, linear, network effects functions found usually in the literature. There are several reasons to choose this more complex form as opposed to a simpler linear one. First, this function is founded on an explicit derivation from a particular model of competition in the software market. Second, the linear form, at least when coupled with a linear heterogeneity function (as is usually the case), severely limits the form and set of possible equilibria – most obviously there is only one interior equilibrium configuration and this is necessarily symmetric (if network effects and heterogeneity are symmetric, i.e. a priori the platforms are equivalent). Third, and most importantly, as we discuss in detail below, the form of the network effects function is a key determinant of comparative statics for welfare. In this regard, the use of the linear form is not 'innocent' and has strong implications for the results. While this is obviously true of any other form chosen, including the one here, examining a slightly more complex, and less standard case, forces us to think more carefully about the implications of choosing one particular functional form over another.

Coming to the heterogeneity function it is set to take the form $h_A(t) = 10t^{10}$. This corresponds to a situation where there is a large middle ground of consumers who are fairly indifferent between the two platforms (h(t) is small until t is close to 1) but two 'extreme' groups at either end who have strong preferences for their nearest platform. The high power (t^{10}) was determined by the need to ensure the existence of a stable asymmetric equilibrium and itself reflects the sharp concavity of the network effects function. Fixed costs are set as follows $f_B = f_A = 1.5$. These values are chosen so as to generate a stable asymmetric equilibrium as shown in Figure 6.1. Note that in its general shape (i.e. number of equilibria, location of maxima/minima) this graph is the simplest possible that gives rise to a stable asymmetric equilibrium.[20]

[19] It has already been assumed above that the network effects are symmetric, that is $\nu_A = \nu_B$.

[20] To have an interior stable equilibrium $A(t)$ must intersect the line $y = 0$ from above. If heterogeneity is symmetric, $h_A(t) = h_B(1-t) = h(t)$ then when fixed costs are equal and prices are zero, $A(t)$ must be anti-symmetric about 0.5, i.e. $A(t) = -A(1-t)$. This implies $A(0.5) = 0$ so 0.5 is an equilibrium. Thus with symmetry in the network function and assuming that standardization equilibria exist (i.e. 0 and 1 are equilibrium) the fewest crossings (i.e. interior equilibria) that lead to the existence of a stable asymmetric equilibrium is five and we must have a situation similar to that shown.

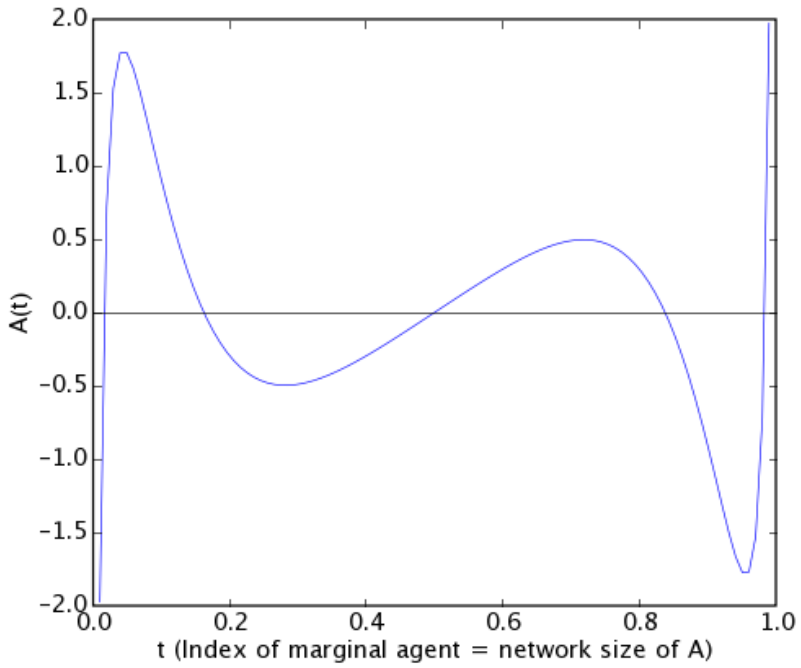

Figure 6.1: The utility advantage function, $A(t)$ in the symmetric case when the access prices for the two platforms are the same (so $p_A = 0$) and $f_A = f_B = 1.5$. There are stable equilibria at 0 and 1 (the 'standardization' equilibria) and 0.16 and 0.84 (asymmetric stable equilibria). There are unstable equilibria at 0.5 and 0.02 and 0.98.

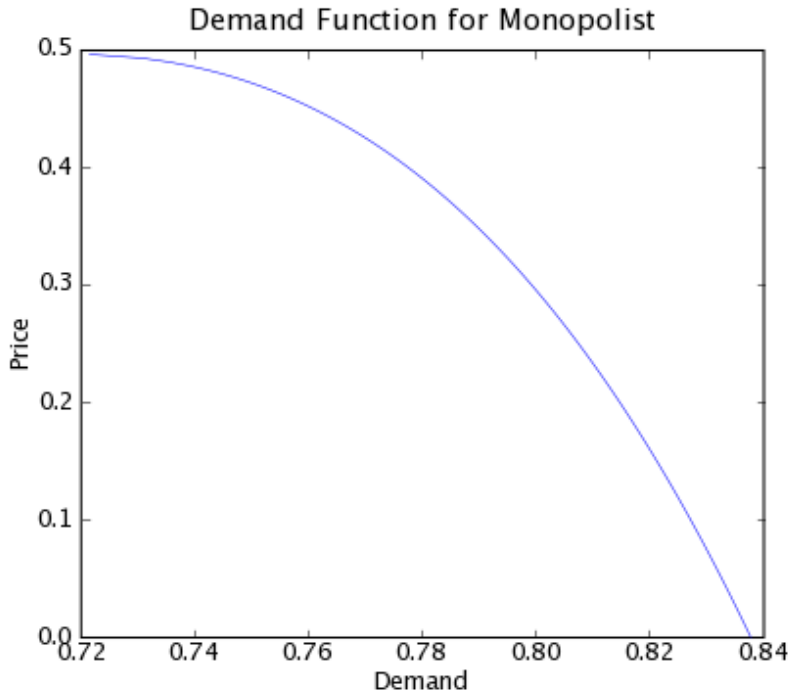

Figure 6.2: The Demand function for the monopolist in the neighbourhood of the stable equilibrium at 0.84. Demand is discontinuous at a price just below 0.5 (i.e. at the left edge of the diagram – the discontinuity itself is not shown as it distorts the scale).

Discontinuity of demand:

Since price enters $A(t)$ linearly the diagram above also implicitly defines the demand function in the neighbourhood of an equilibrium (an increase in the p_A shifts the $A(t)$ curve down by that amount). A maximum of $A(t)$ therefore corresponds to a point at which demand is discontinuous (as price rises above the maximum value demand jumps down as the market tips to the neighbourhood of next lowest stable equilibrium).

An illustration of this is provided in Figure 6.2, which plots the demand function derived from Figure 6.1 in the neighbourhood of the stable equilibrium at 0.84. Here demand is discontinuous at a price just below 0.5 (i.e. at the left edge of the diagram – the discontinuity itself is not shown as it distorts the scale). At the discontinuity demand will suddenly jump down to approximately 0.14 which is the next place the line y=0.5 would intersect $A(t)$ (see Figure 6.1). Note how this diagram is just the relevant portion of Figure 6.1 between 0.73 and 0.84 'blown up'.

In all cases where there is symmetry and a stable asymmetric equilibrium $A(t)$ must have a bounded maximum just like it does in Figure 6.1. A bounded maximum in turn implies a discontinuity in the demand function of the monopolist. Thus, in all such cases, *a monopolist will face a discontinuous demand function*. This discontinuity in

demand does not exist in the traditional linear network effects models and it functions here to place a sharp upper bound on the price the monopolist can charge without a sudden jump downwards in market share.

Other Comparative Statics:

We can evaluate the effect of changing production and porting costs by considering how it shifts $A(t)$. In particular, increasing fixed costs of software production for A f_A will shift $A(t)$ down and increasing f_B will have the opposite effect (note that f_B is equal to the porting costs, f^p if the porting constraint does not bind). Note that unlike price, fixed costs do not enter linearly so they will also change the shape of $A(t)$ and not just its level.

6.4 Welfare

Having established the various properties of equilibrium in this section we come to the central questions of this paper: how does the monopolist's control of prices and the cost of porting affect consumer and social welfare? Giving equal weight to monopoly profits and consumer welfare we have that total welfare, $W = \Pi_A + W^C$ where W^C is consumer welfare and Π_A are the monopolist's profits.[21]

Lemma 6.4.1. *The marginal change in consumer welfare as a function of platform A's size (t) is:*

$$\frac{dW^C}{dt} = A(t) + \mu(t)$$

where $A(t)$ is the utility advantage of A over B defined previously and

$$\mu(t) \equiv t\nu'_A(t) - (1-t)\nu'_B(1-t)$$

At an interior equilibrium $t_e \in (0,1), A(t_e) = 0$, and this reduces to:

$$\frac{dW^C}{dt} = \mu(t_e)$$

Proof. See appendix. □

A first point to emphasize is that this result (and Lemma 6.4.2 below) are entirely general and will hold in *any* model in which consumer utility incorporates a 'network effects' function (whether arising directly, or, indirectly as a reduced form derived from

[21] We have assumed overall profits are zero in the software industry as a result of free entry.

a more complex model). That is, there is nothing that depends on the specifics of the porting framework as presented in this paper. In particular, these results would apply both traditional direct network effects models of communication networks and some of the more recent models arising from a two-sided market structure.

The second point to make is that this result is telling us that, when at an interior equilibrium ($x = t_e$), the marginal change in consumer welfare with respect to platform size is *a function of 'network effects' alone* (encapsulated in μ). The two basic possibilities, namely that consumer welfare is increasing ($\mu(t_e) > 0$) or that it is decreasing ($\mu(t_e) < 0$) with the size of platform A have a simple interpretation. In the first case we have a situation in which more standardization (that is more consumers on platform A) is preferable. In the second case we have a situation in which more symmetrical platform shares are preferable.[22]

What determines the sign of $\mu(t_e)$ – that is whether standardization or symmetry is preferable? Answer: the degree of curvature of the indirect network effects function, ν, which in more economic terminology could be put as: how sharp are the diminishing returns to network effects in platform size, that is, how fast does the benefit of a new user fall as the number of users on the platform increases – crudely how much (less) does an existing user of a platform benefit as the millionth person joins a platform compared to when the tenth person joins?).

Interestingly it turns out that the dividing line between the two cases is where network effects take the form of the natural logarithm: $\nu(x) = C + ln(x)$. When marginal network effects fall with platform size more gradually than this then $\mu > 0$ and standardization is preferable. When marginal network effects fall more strongly than this then $\mu < 0$ and symmetry is preferable. The classic form studied in the literature is of course where ν is linear in which case marginal network effects do not fall at all with platform size and so $\mu > 0$ and standardization is preferable. Conversely, the circular city model of indirect network effects studied in the appendix gives rise to the case where $\nu(x) \propto -1/\sqrt{x}$. In this case marginal network effects fall more sharply than for the logarithm and so $\mu < 0$.[23]

To summarize, network effects which display weakly diminishing returns imply that standardization (everyone on one platform) will be preferable while if network effects

[22]There is, also the third possibility that the change in consumer welfare is zero but this is obviously a very special case.

[23]Odlyzko and Tilly (2005) provide a thoughtful critique of existing assumptions regarding the form of the network effects function such as that embodied in Metcalfe's law (Metcalfe's law corresponds to the linear case $\nu(x) = x$). Interestingly, as a replacement they propose using the logarithmic form, $\nu(x) = ln(x)$. As we have just shown this is a very special case in which at an equilibrium we have $\mu = 0$ and therefore consumer welfare is neither increasing or decreasing in platform size. Clearly, one would like to determine the exact form of the (indirect) platform effects function empirically. However, at least to our knowledge, there are no *economic* papers which deal with this issue.

show strongly diminishing returns, a more symmetric platform configuration is preferable. We now proceed to work formally through the consequences of this basic result in relation to the model at hand in the Lemmas below, with the main results summarized in Table 6.1.

Lemma 6.4.2. *At a subgame equilibrium, t_e, the effect on consumer welfare of an increase in the price charged by the monopolist is negative if $\mu(t_e) \geq 0$ and is ambiguous otherwise depending on the relative magnitudes of the monopoly pricing effect (-ve) and the network externality (+ve). Furthermore, at an equilibrium of the overall game (i.e. where the monopolist is profit-maximizing) the change in total welfare equals that in consumer welfare and therefore has the same properties.*

Proof. See appendix. □

Monopoly pricing does not result in traditional deadweight losses since total demand is fixed and does not change (consumers who leave one platform join the other).[24] However, it does shift consumers away from the monopolist's platform (an effect exacerbated by the feedback from the indirect network effects). In markets with 'externalities' such as these this will have consequences for welfare.

The effect of an increase in the monopolist's price depends on two distinct factors. The first factor is the simple one that higher prices reduce consumer welfare because consumers pay more. The second factor is more subtle. An increase in M's price moves consumers off A onto B. This effect may either be negative or positive depending, respectively, on whether a more standardization-type or a more symmetric platform configuration is better for welfare. As shown in Lemma 6.4.1 this second condition is equivalent to asking whether $\mu(t_e)$ is positive (standardization-type better) or negative (symmetric better). Thus, if $\mu(t_e)$ is positive, an increase in the monopoly price by reducing the size of platform A acts to reduce welfare. Conversely when more symmetric platform sizes are preferred then an increase in the monopoly price by reducing the size of platform A actually acts to increase welfare.

If we combine the two factors then we only get an unambiguous prediction (increase in prices reduces welfare) in the first case, that is when a more standardization-type platform configuration is preferable. In the second case, where a more symmetric platform configuration is preferable, the effect will be ambiguous and welfare could actually rise due to an increase in the monopolist's prices.

Lemma 6.4.3. *At a subgame equilibrium, t_e, the effect on consumer welfare of a increase in porting costs is negative if $\mu(t_e) < 0$ and is ambiguous otherwise depending on*

[24] This explains why at full equilibrium marginal consumer welfare and total welfare are equal.

	Low Curvature	**High Curvature**
Direct Impact of Higher Price	-	-
Indirect Impact of Higher Price	-	+
Overall Impact of Higher Price	-	O
Direct Impact of Higher Porting Cost	-	-
Indirect Impact of Higher Porting Cost	+	-
Overall Impact of Higher Porting Cost	O	-

Table 6.1: Welfare Impact of Changes in Price and Porting Cost. This table summarizes the results of Lemmas 6.4.2 and 6.4.3. Curvature refers to the curvature of the network effects function in the neighbourhood of an equilibrium (note that at an equilibrium consumer and social welfare are equal). 'O' indicates the effect is ambiguous.

the relative magnitudes of the welfare loss from a direct reduction in software provision on platform B and the welfare gain from an increase in A's market share. Furthermore, at an equilibrium of the overall game (i.e. where the monopolist is profit-maximizing) the marginal effect on total welfare equals the marginal effect on consumer welfare.

Proof. See appendix. □

Again we have two distinct effects of higher porting costs. The first, and the direct one, is that higher porting costs result in a reduction in availability of software for those on platform B (and probably higher prices too – though this may depend on the specifics of the model for software provision). This unambiguously reduces welfare because higher porting costs mean less software for B users (holding platform B's share constant).

The second effect arises from the fact that, as a result of the change in software availability on B, some consumers move from platform B to platform A. This change is an exactly similar one to that already analyzed above when discussing the effect of a price rise (except here an increase porting cost increases the size of platform A while an increase in price reduces the size of platform A). In particular the effect will be negative if, and only if, $\mu(t_e)$ is negative (more symmetric platform configuration preferred). In this case, both effects operate in the same direction and an increase in porting cost is unambiguously harmful to consumer welfare. On the other hand if a more standardization-type platform is preferable ($\mu(t_e) > 0$) then this effect is positive and the overall impact on welfare will depend on the relative magnitude of the two effects. In this second ambiguous case, we can explore the 'second order' comparative statics in more detail, and this is done in the next Lemma.

Lemma 6.4.4. *At a subgame equilibrium, t_e, if $\mu(t^e) \geq 0$ so that the effect of porting costs on consumer welfare is ambiguous, then it is more likely that the effect is negative:*

- *The larger is platform B's market share (more consumers to suffer from the reduction on software provision on B)*
- *The larger is the direct impact of higher porting costs on the provision of software for B (greater reduction on software provision on B).*
- *The smaller is the impact of changes in porting cost on A's market share.*[25]
- *The smaller is the increase in consumer welfare of an increase in A's market share.*

Proof. See appendix. □

6.5 Example II: Welfare

We now return to our previous specific example, this time in order to illustrate the welfare analysis. Using it, among other results, we demonstrate that it is possible for the welfare costs (consumer or societal) of the control of porting to be significantly greater than the costs of monopoly pricing.

We first choose specific functional forms and values for constants. The heterogeneity function is chosen to ensure that there exists an asymmetric stable equilibrium and is the same as that used for figure 2 above: $h(t) = 10t^{10}$.

The direct costs of software production are set to $f_A = 1.5$ and the initial porting cost is set to two-thirds of that value, so $f^p = 1.0$. The monopolist's expenditure function is: $e(f^p) = 2 \cdot (f^p - 1)^4$ and the initial value of f^p when there are no efforts by the monopolist is set to 1. The expenditure function displays diminishing returns and while initial efforts to prevent porting are relatively cheap the cost then escalates rapidly.

The exact parameters for the functional form of the expenditure function are chosen so that an interior 'porting cost' solution exists i.e. the value of porting cost obtained is such that $f_A > f^p$ and expenditure to prevent porting is non-zero and non-infinite. Using these values we can now proceed to solve the monopolist's problem by numerical means and have the following results.

We find the values chosen for the two control variables are 1.419 for porting costs and 0.43 for the access price of platform A. We also calculate the profit-maximizing price M would charge when unable to influence porting costs: 0.079. Our main interest

[25]For example, if the main effect of changes in porting cost were to soften competition rather than to directly increase A's market share. That is, in terms of A's demand curve, increasing porting costs steepened the demand curve or shifted it up but did not shift it out.

	Porting Cost	Price of A Hard-ware	A's market share	Net Profits for M	Consumer Welfare	Total Wel-fare
Initial porting cost, competitive prices	1.0	0	0.758	0	0.0	0.0
Initial porting cost, monopoly price on A	1.0	0.079	0.704	0.056	-0.046	0.010
Monopolist chosen porting cost, monopoly price on A	1.419	0.43	0.729	0.252	-0.406	-0.154

Table 6.2: Welfare Results at Various Prices and Porting Costs

is in the significance of M's choices for welfare and welfare outcomes. These, along with the values of other significant variables, are presented in Table 1 (NB: since ϕ is an arbitrary constant it has been set so that initial welfare values are normalized to zero. This value has no significance since, as already explained, welfare can be changed by a fixed constant (ϕ). Thus only the sizes of welfare *changes* can be meaningfully compared.)

The first line is there to show the baseline case, when the control parameters are at their 'default' values (that is without intervention by the monopolist). In this case, M's market share, with its own price at zero and the fixed costs of porting at 1, is still 75%. Total welfare and consumer welfare are the same – since prices are zero – and has been normalized to zero.

The next line shows the situation if the monopolist can only set prices and is not able to influence porting costs. This helps us benchmark the relative gain to a monopolist of being able to influence porting costs in addition to setting prices. In line with theory the welfare change is slightly positive, reflecting the reduction in the size of Network A.

The final line shows the actual outcome with the porting cost and price at the level chosen by M to maximize its profits. Porting costs increase by almost a half to 1.42, nearly reaching the same levels as the cost of direct production (1.5). Prices rise by over five times compared to the situation when porting costs can not be altered demonstrating the large impact of the Monopolist's control of porting. Despite the far

higher price, market share for the monopolist rises though it is still lower than in the situation where neither price nor porting cost can be set.

6.5.1 The Monopolist's Profits

M gains dramatically from the ability to manipulate porting costs, the percentage increase in profits being approximately 400% over what is obtained when porting costs are fixed. Moreover this is net of the costs incurred to prevent porting, $e(f^p) = 0.0616$, which are equal to a fifth of gross profits. The main effect of raising porting costs is not to increase market share but to soften competition between the two platforms and therefore permit a much higher profit-maximizing price to be charged. Market share at the monopoly price in the two cases when porting cost is and is not manipulatable are quite close (0.704 vs. 0.729).

6.5.2 Consumer welfare

The change in consumer welfare from monopoly pricing, $\Delta W_c^M = -0.046$. The change resulting from higher pricing and higher porting costs is $\Delta W_c^{Mf} = -0.406$. Thus consumer welfare losses arising from the combination of higher porting costs and higher prices are almost *nine times* as large as those arising from higher prices alone.[26]

6.5.3 Total welfare

For total welfare increasing M's price will actually increase welfare: with porting cost at 1, $\Delta W^M = 0.01$. However the welfare change due to the combination of monopoly pricing and higher porting costs is decidedly negative $\Delta W^{Mf} = -0.156$. Thus for this case welfare costs go from barely positive to significantly negative.

6.5.4 Alternative Specifications

This example is of course based on only one set of functional forms and one set of parameter values among many. It is therefore natural ask how specific the estimates presented here are to those particular choices.

[26] As already stated, as welfare is only defined up to a constant we can only compare changes in welfare and not levels. Nevertheless, utility is money metric (prices enter linearly) and profits are well-defined so it *is* possible to convert of welfare *changes* into monetary terms. As a very simple 'back-of-the-envelope' calculation consider applying this analysis to the Microsoft case. Profits in 2000 (around the time of the antitrust settlement in the US) were approximately \$9.5 billion and in our model profits equal 0.252. Thus, in dollar terms the change in consumer welfare from monopoly pricing alone equals approximately \$1.7 billion ($0.046/0.252 \cdot 9.5$), while the change in consumer welfare with both higher prices and higher porting costs equals \$15.3 billion ($0.406/0.252 \cdot 9.5$).

In many ways we are rather limited in what we can say: those general results that are obtainable have already been presented in the previous section. As shown there the welfare impact of a change in price and porting cost depend crucially on the rate of diminishing returns of the network effects function. With strongly diminishing returns pricing has an ambiguous impact but porting costs have a negative impact but with weakly diminishing returns we have the converse: a negative price impact and an ambiguous impact of porting cost.

Thus the choice of network effects function to use in a simulation will clearly influence the estimated welfare impact. The example here uses an indirect network effects function which displays strongly diminishing returns – and consistent with the general results we find a weakly positive impact of pricing and a negative impact of porting cost. However if one were to use a network effects function with weakly diminishing returns (for example linear network effects) this would likely change the results – it would certainly make it more likely that the pricing impact on welfare was more significant than the porting impact.[27]

6.6 Conclusion

In this paper, we introduced 'porting' into a standard, two-sided, indirect network effects model, with 'porting' playing a role analogous to 'converters' in the simpler direct network effects models. With 'porting', software developed for one platform can be converted to run on another (at a cost lower than that of direct production). We examined general properties of this model, looking, in particular, at what occurs when one (dominant) platform is controlled by a single firm, the Monopolist, who is able to control the cost of porting to a competitor platform (at the cost of some expenditure on the Monopolist's part). We demonstrated the existence of a platform (and porting) equilibrium and examined various associated properties, such as the discontinuity in the monopolist's demand function.

Next we turned to the question of consumer and social welfare. It was shown that, the effect on welfare both of monopoly pricing and higher porting costs depended crucially on the degree of diminishing returns to platform size in the indirect network effects function (ν). If diminishing returns were weak then monopoly pricing had a negative effect on welfare but the effect of the higher porting costs was ambiguous, while with strongly diminishing returns the converse held, that is the effect of monopoly pricing was ambiguous but higher porting costs had a negative effect.

[27] In fact it would no doubt be possible to choose a model such that the ability to control porting increased welfare – all one would need is for the benefits of platform 'standardization' to be sufficiently strong.

Finally, we provided an illustrative example using a specific case of our model. We showed that, in this example, the social and consumer welfare losses arising from the control of porting combined with monopoly pricing dwarfed the welfare effects stemming from monopoly pricing alone. In particular, consumer welfare losses from the combination of higher porting costs and higher prices were over nine times higher than those arising from higher prices alone. For total welfare, there was almost no effect of monopoly pricing alone but a significant reduction when the monopolist controlled both prices and porting costs (in this second case the welfare loss was equal to approximately three fifths of the monopolist's profits). Of course this is a single example and without either calibrating from empirical data or extensive robustness-checking one would not wish to use the results for policy-making. Nevertheless, it does provide a useful example that helps put flesh on the dry bones of the general model.

These results, taken together, have important consequences for competition policy. They demonstrate how, in a two-sided market environment, anti-competitive behaviour may manifest indirectly through actions taken to control porting rather than through direct tying or pricing behaviour. Furthermore, for the monopolist the benefits of controlling porting may also accrue indirectly: that is, by increasing the prices that can be charged at a given level of demand rather than increasing demand. Returning to the examples discussed in the introduction, we would suggest that an analysis based on the control of porting provides a better way of understanding the effects and motivations of a dominant firm than alternative approaches, such as those based on traditional theories of tying or even switching costs.[28]

Of course from an antitrust point of view this is not enough – simply establishing a potential 'anti-competitive' motivation for a firms behaviour is not sufficient to show such actions will actually harm welfare. In this regard, as already mentioned, our central result was that the crucial parameter to estimate is the curvature of the indirect network effects function (that is the degree of diminishing returns to platform size). When the degree of diminishing returns is high – the benefit of a millionth user is much less than the thousandth – the control of porting unambiguously harms welfare but when the degree of diminishing returns is low – the benefit of the millionth user and the thousandth user is similar – then the control of porting has an ambiguous impact (it may even increase welfare). Given this, the first step for an antitrust economist tasked with analyzing the control of porting in a particular industry would be to estimate the form of the indirect network effects function for the particular platforms under

[28]Though, of course, in one sense the control of porting can be seen as a special case of tying (or the creation of switching cost) in which the 'tie' is not aimed at competing providers of the tied good but at the owners of competing *platforms*.

consideration.[29]

When the control of porting does harm welfare, policy-makers may wish to take steps to reduce the control of porting by a dominant firm. One simple way to do this is to promote 'open standards' at the interface between the 'software/service' and the platform. For example, in the case of TPMs/DRMs (Technological Protection Measures/Digital Rights Management) systems a policy-maker could promote (or require) interoperability between different TPM/DRM systems so that the music ('software' in our terminology) purchased from any given vendor will work on any given digital music player (the platform).[30]

Similarly, in the case of the EU dispute with Microsoft over Microsoft's Windows Media Player, rather than requiring unbundling the authorities could simply require that any audio formats specific to Windows Media Player must be 'open' and freely licensable so as to ensure that it is easy to port music and complementary services to a media player on another platform such as Linux. The same approach would also apply to web browsers where there already exist an extensive set of open standards developed by the W3C. Again, rather than requiring Microsoft to unbundle Internet Explorer the authorities could simply press for 'standards-compatibility'. In this way developers of websites and other forms of web-services would be able to develop in a platform-neutral way (essentially the cost of porting to a different platform such as Linux+Firefox would then be zero) with all the associated long-run benefits for competition and consumer choice.

Finally, we mention some of potential avenues for future work. One of the most obvious improvements that could be made would be to replace the simple monopoly model with an oligopoly in which each platform has a profit-maximizing owner. Porting, and the manner in which it may be controlled, have been modelled in a fairly simple manner. One might improve this in various ways. For example, one could change from a 'black box' cost function e to a setup where f_A increases with f^p – this would correspond to an 'obfuscation' situation where increasing porting costs to competitor platforms also increases the cost of producing software on one's own platform.

One could also add dynamics to the model (though this would also greatly increase complexity). For example, rather than having a fixed static demand one could allow consumers to arrive over time.[31] Alternatively consumers could make repeat purchases

[29]As mentioned in an earlier footnote there is little empirical evidence for the form of 'network effects' (whether in two-sided markets or traditional 'network' industries). The fact that most of the models in the theoretical literature use a linear specification is due solely to analytical tractability and not to any empirical support for this functional form – a choice which, as this paper has shown, is not an innocent one.

[30]At the present time this very issue of DRM interoperability is being debated both at the EU level and in various individual European countries in relation to Apple's FairPlay DRM.

[31]See, for example, Cabral (2007) or the model of Fudenberg and Tirole (2000) which gives rise to

but with a switching cost if a different platform were chosen in a subsequent period.

Finally, it would be interesting to explore the consequences of allowing for innovation in software provision perhaps via the introduction of a quality ladder. Such an approach would raise additional thorny questions about the welfare impact of monopolist behaviour if innovation were not barrier to entry neutral. For example, if innovations while increasing quality also made it easier to port from one platform to another (consider the case of Java or the emergence of the web and web browsers as a fully-fledged application development platform).[32] In this case, efforts to obstruct porting would also hinder innovation, with all the attendant consequences for welfare.

6.A Proofs

6.A.1 Proof of Lemma 6.3.1

Recall that the *conditional utility advantage* of platform A over platform B for consumer t when platform size is n_A:

$$\hat{A}(t, n_A) = u_A(t, n_A) - u_B(t, 1 - n_A)$$

and the *utility advantage (function)*, which gives the utility advantage of platform A over B if t is the marginal consumer (so $t = n_A$):

$$A(t) = \hat{A}(t, t)$$

Suppressing n_A for the time being we shall simply write $\hat{A}(t)$.

Since 'heterogeneity cost' for a consumer is increasing in the distance of the consumer from the chosen platform we have that $\forall t, \hat{A}'(t) < 0$. Then $\hat{A}(t_m) > 0$ implies $\hat{A}(t) > 0, \forall t \leq t_m$. Conversely if $\hat{A}(t_m) < 0$ then $\hat{A}(t) < 0 \forall t \geq t_m$.

Now a consumer (with expectations of platform A size equal to n_A) chooses platform A over B iff $\hat{A}(t) \geq 0$. Thus if a consumer with index t_m chooses platform A then all consumers with index $t \in [0, t_m]$ choose platform A. Similarly if a consumer with index t_m chooses platform B then all consumers with index $t \in (t_m, 1]$ choose platform B.

In particular this immediately implies that if there exists $t_m \in [0, 1], \hat{A}(t_m) = 0$ (and there is at most one such solution since $\hat{A}' < 0$) then this is the marginal consumer and the resulting platform size of A is t_m. This is because for $t \in [0, t_m], \hat{A}(t) > 0$ so

limit-pricing behaviour on the part of the monopolist. Though we note that the addition of dynamics adds very substantial technical complexity.

[32] See e.g. Farrell and Katz (2000) on network monopolies and downstream innovation.

these consumers choose platform A while for $t \in (t_m, 1], \hat{A}(t) < 0$ so these consumers choose platform B.

For the extremal cases by the same arguments if $\hat{A}(0) < 0$ then all consumers choose platform B and if $\hat{A}(1) > 0$ then all consumer's choose platform A.

Furthermore, only one of these alternatives is possible so there is a unique implied platform size for any given assumed n_A. Thus one may define a function $f : [0,1] \to [0,1]$ where for a given assumed platform size, $n, f(n)$ is the resulting implied platform size.

Imposing rational expectations then implies that n_A is an equilibrium if and only if n_A is a fixed point of f. But n_A is a solution of $f(n) = n \Leftrightarrow n_A \in E$. QED

Remark: Equilibria $t \in E_{-0}$ are often termed standardization or tipping equilibria as they involve all consumers joining a single platform.

Remark: This result sets up an implicit equivalence between platform size and the marginal consumer (where the term marginal is broadened to include the tipping situations where $t_m = 0$ or 1 and $A(t_m) \neq 0$

Stability of Equilibria: Suppose we have equilibrium $t_m \in E_0$ with $A'(t_m) < 0$. Suppose that there is a perturbation in expectations so that a platform size of $t_m + \epsilon$ is expected instead of t_m (where $\epsilon > 0$). Since $A' < 0$ we must have $\hat{A}(t_m + \epsilon, t_m + \epsilon) = A(t_m + \epsilon) < 0$. Now in the interior all functions are continuous so $\hat{A}$ is continuous. Thus δ in the region $t_m + \epsilon$ we have that $\hat{A}(x, t_m + \epsilon) < 0$ for $x \in (t_m + \epsilon - \delta, t_m + \epsilon]$. But then all consumers with indices in that range wish to leave platform A and go to platform B. Repeating this process we converge back to the equilibrium t_m. The analogous argument for negative ϵ shows the equilibrium is stable to perturbation downwards in expectations. Thus the equilibrium is stable.

The exact same form of argument applied to an equilibrium $t_m \in E_{-0}$ shows that it too is stable. QED.

6.A.2 Proof of Lemma 6.3.2 (Porting Lemma)

The result will follow from two claims:
Claim 1: Suppose that a platform has a piece of software produced directly for it. Then s_X, p_X^s are determined by f_X^d (the direct cost of software production) alone. We may therefore take $f_X = f_X^d$ in all the formulas obtained above (it is immaterial for the purposes of calculating all equilibrium values whether software is ported or produced directly for this platform).

Proof. The cost of porting is less than the cost of direct production. Thus as long as one software firm enters directly it must be the profit condition of that firm that binds

(i.e. is zero). This condition alone determines the total number of software firms and software prices. □

Clearly if no firm produces directly there can be no porting as there would be nothing to port.

Claim 2: If porting is possible in both directions and both platforms have some software produced directly then both platforms have the same amount of software produced for them. This in turn implies $f_A^d n_B = n_A f_B^d$.

Proof. If software is produced directly then all software that could have ported must have been (since it is cheaper to port). Let d, p (d', p') be the amount of directly produced software and ported software respectively on A (B). Then $s_A = d + p$ but $p' = d, p = d'$ so $s_A = s_B$. If this is the case it requires $f_A^d n_B = n_A f_B^d$ since $s_X^2 f_X^d = n_X$. □

The statement of the Lemma specifically excludes the possibility that $f_A^d n_B = n_A f_B^d$. This immediately implies the converse of the claim, namely that that software is produced directly for at most one platform. The Lemma is proved. QED.

6.A.3 Proof of Lemma 6.3.4

Proof of existence: Fix an equilibrium $t_e^0 \in E_0(p_A^0, ...)$ then we can define $t_e(p_A, ...)$ by picking $t_e \in E(p_A, ...)$ consistent with t_e^0. Since $A(t)$ is continuously differentiable so too will be $t_e(p_A, ...)$ (at least almost everywhere – see below). For notational convenience whenever a parameter is fixed we shall drop it from the list of arguments to $t, A, ...$

Differentials: implicitly differentiate the equation $A(t) = 0$ with respect to the relevant variable (p_A, f_A, f_B). Since increasing A's price by dp shifts the $A(t)$ curve down by dp reducing t_e the sign of the differential is as stated. Similarly increasing f_A shifts the platform advantage curve down and therefore the advantage curve down reducing t_e and therefore the differential with respect to f_A must be negative (and conversely for f_B).

Remarks on discontinuity and profit maximization: Fix f_A, f_B, then $t_e(p_A) = A^{-1}(p_A)$ is the demand function faced by M. From the previous result we know this is downward sloping. Now take a stable equilibrium t^0 when $p_A = 0$ and assume there exists an adjacent non-extremal equilibrium $t^{0'} \leq t^0$ (which must be unstable). Then there must exist a maximum of $A(t)$ at $t^1 \in (t^{0'}, t^0)$ with $A'(t^1) = 0$ and the demand function $t_e(p_A)(t_e(0) = t^0)$ is discontinuous at t^1 with $p_A^d = A(t^1)$.

Despite this there will still exist a profit maximizing price $p_A^d > p_A^m$ since

$$\lim_{t \to t_+^1} A^{-1}(t) = -\infty$$

6.A.4 Proof of Lemma 6.3.5

Set all of M's control variables to their initial value. Suppose first there are no discontinuities in M's demand function. This occurs iff there exists no zeroes of $A'(t)$, i.e iff $A(t)$ is monotonic. Since we assume existence of a stable interior equilibria must have that $A(t)$ is downward sloping. Thus we have a downward sloping demand function. This gives a well-defined and continuous profit function on a compact set (demand space extends only from 0 to 1). Thus the profit function has a maximum which it attains somewhere on the set. QED.

Again, set all of M's control variables to their initial value. So assume that there there is a discontinuity in the demand function, i.e. that there exists a t with $A'(t) = 0$. Pick an interior stable equilibrium. Then by Lemma 6.3.4 there exists an associated well-defined demand function. Furthermore, there exists locally a unique profit-maximizing price which occurs prior to any discontinuous jump (downwards) in the demand function. But this ensures the existence of equilibrium in the overall game since it means that at any discontinuity in the demand function the profit function is downwards sloping. QED.

6.A.5 Proof of Welfare-Related Propositions

Consumer welfare as a function of platform A's size (t) is given by (for simplicity ϕ is omitted):

$$W^C(t) = -t \cdot p_A + t\nu_A(t) + (1-t)\nu_B(1-t) - \int_0^t h_A(x)dx - \int_t^1 h_B(x)dx$$

Moving to total welfare we need only add in the relevant expression for $\Pi_A = t \cdot p_A - e(f^p)$. Thus:

$$W = t \cdot p_A - e(f^p) - t \cdot p_A + t\nu_A(t) + (1-t)\nu_b(1-t) - \int_0^t h_A(t)dt - \int_t^1 h_B(t)dt$$

Proof of Lemma 6.4.1

Differentiating consumer welfare with respect to t yields:

$$\frac{dW^C}{dt} = -p_A + \nu_A(t) - \nu_B(1-t) - h_A(t) + h_B(1-t) + t\nu'_A(t) - (1-t)\nu'_A(1-t)$$

This simplifies to ($A(t)$ is the utility advantage of A over B defined previously):

$$\frac{dW^C}{dt} = A(t) + t\nu'_A(t) - (1-t)\nu'_B(1-t) = A(t) + \mu(t)$$

where we have defined:

$$\mu(t) = t\nu'_A(t) - (1-t)\nu'_B(1-t)$$

At an equilibrium $t_e, A(t_e) = 0$, so this reduces to:

$$\frac{dW^C}{dx} = \mu(t_e)$$

QED.

Proof of Lemma 6.4.2

$$\frac{dW^C}{dp_A} = -t + \frac{dt}{dp_A}\frac{dW^C}{dt}$$

Considered at an asymmetric equilibrium the second term will be greater than or less than zero depending on whether μ is less than or greater than zero. If μ is non-negative then the second term is negative and total sum will be negative. If μ is negative the total sum will be ambiguous (depending on the relative magnitudes of the two terms). Thus, if network effects do not show very strong diminishing returns (and so μ is non-negative) welfare changes negatively with increasing price. If μ is negative (as it would in the circular city model) then the effect on consumer welfare depends on the relative size of the monopoly pricing costs (first term) versus the network externality (second term).

Turning to total welfare we have:

$$\frac{dW}{dp_A} = \frac{d\Pi_A}{dp_A} + \frac{dW^C}{dp_A} = \frac{dt}{dp_A}(p_A + \frac{dW^C}{dt})$$

The term outside the brackets is negative but again here the second term can have either positive or negative sign in general. NB: when the monopolist is profit maximizing the differential of monopolist profits with respect to price is zero. Thus, the differential of total welfare equals the differential of consumer welfare.

Proof of Lemma 6.4.3

The change in consumer welfare as a consequence of an increase in the cost of porting is:

$$\frac{dW^C}{df^p} = (1-t)\frac{d\nu_B}{df^p} + \frac{dt}{df^p}\frac{dW^C}{dt}$$

The first term is clearly negative since software provision on platform B declines as porting costs go up. The analysis of the second term is similar to the case of a change in price. As platform A's market share increases as porting costs increase the second term will be greater than or less than zero depending on whether μ is greater than or less than zero. Thus, if μ is less than zero (strongly diminishing marginal network effects) the total will be unambiguously negative and consumer welfare declines with increases in porting costs. If μ is positive then the total has ambiguous sign in general, and will depend on relative sizes of the two terms.

For total welfare we have:

$$\frac{dW}{df^p} = \frac{d\Pi_A}{df^p} + \frac{dW^C}{df^p}$$

When profit-maximizing the first term is zero and the differential of total welfare equals that of consumer welfare. When not at a profit-maximizing level of porting costs the first term is positive. In this case whether the total is positive or negative will depend on the specific circumstances.

6.A.6 Proof of Lemma 6.4.4

The stated results all follow by straightforward examination of changes in the sizes of the various terms in the proof of Lemma 6.4.3 above.

6.B Software Production

There are two main methods of modelling product variety in the literature. One based on monopolistic competition and one based on locational models. The monopolistic competition approach has already been extensively used to demonstrate indirect network effects in hardware/software systems (see e.g. Church and Gandal (1992); Church, Gandal, and Krause (2003)). One can also use an approach based on locational differentiation and that is the approach we adopt here.

Software firms on platform X have fixed costs f_X and marginal costs c_X^s. Marginal costs are assumed to be constant across the two platforms but fixed costs are not. Because software production involves a fixed cost it cannot be provided competitively. Instead we introduce a locational model of product differentiation and imperfect competition

For each platform, software 'space' is represented as a circle (of circumference 1). Software firms are assumed to locate symmetrically (and therefore equidistantly) in this space.[33] while consumers are distributed uniformly over it (so total demand for software on platform X is the total number of consumers on that platform: n_X). Following the standard circular city model[34] we have consumer's (expected) utility from software consumption is:

$$u_X^s(s_X, p_X^s) = -E[d(x(s_X))] - p_X^s$$

Where d is a 'travel' cost function of all locational models, $x(s_X)$ is the distance a consumer is from the nearest software, and E is the expectation operator. Average travel cost is used because it is assumed that consumers make their decision when they do not yet know their exact position in software space relative to software producers. Thus they base their decisions on expected costs (which will be common across consumers). We shall assume a linear travel cost, $d(x) = kx$.

6.B.1 Solving

The main result can be stated in the form of a lemma:

Lemma 6.B.1. *Given expected platform sizes n_X^e the equilibrium level of software production, associated prices, and software utility are:*

$$s_X = \sqrt{\frac{kn_X^e}{f_X}}$$

$$p_X = c_X^s + \sqrt{\frac{kf_X}{n_X^e}}$$

$$u_X^s(s_X, p_X^s) = -c_X^s - \frac{5}{4}\sqrt{\frac{kf_X}{n_X^e}}$$

Proof. The setup is exactly the same as the textbook circular city model (see e.g. Tirole 1988) except that demand rather than being 1 is equal to the expected market size of that platform: n_X^e. This leaves prices unchanged (since the shape of demand curve is unchanged), so in equilibrium: $p_X = c_X^s + \frac{k}{s_X}$ where k is the cost of travel ($d(x) = kx$). Firms locate equidistantly and each face the same level of demand equal

[33]Firms' location decisions could be endogenized and this outcome derived as an equilibrium configuration – see Economides (1989) However we choose to take this as an assumption for the sake of simplicity.

[34]See e.g. Tirole (1988) for details.

to total demand divided by the number of software firms. To determine the number of software firms we use the free entry condition which means that in equilibrium firms earn zero net profits – i.e. they cover fixed costs:

$$(p_X - c_X^s)\frac{n_X^e}{s_X} - f = 0 \Rightarrow \frac{kn_X^e}{s_X^2} - f = 0 \Rightarrow s_X = \sqrt{\frac{kn_X^e}{f}}$$

This in turn gives:

$$p_X = c_X^s + \sqrt{\frac{kf}{n_X^e}}$$

The form of the software utility functions in our particular case? Consumers do not know the exact location of firms in advance so they base their decisions on the expected distance from a software producer. Software firms locate randomly but equidistantly on the circle and consumers are uniformly distributed thus expected distance between a consumer and the nearest software is a quarter of the distance between firms. Distance between firms is the inverse of the number of firms, s_X. We therefore have:

$$u_X^s(s_X, p_X^s) = -p_X^s - k(\frac{1}{4s_X})$$

Substituting the values for p_X, s_X we have[35]:

$$u_X^s(s_X, p_X^s) = -c_X^s - \frac{5}{4}\sqrt{\frac{kf}{n_X^e}}$$

□

Remark: Since the constant $\frac{5\sqrt{k}}{4}$ can be absorbed into fixed cost f_X this variable will be omitted in future and we have:

$$u_X^s(s_X, p_X^s) = -c_X^s - \sqrt{\frac{f}{n_X^e}}$$

We can now substitute this expression for u_X^s to obtain:

Corollary 6.B.2. *The reduced form of the utility function is:*

$$u_X(t) = \phi - p - h_X(t) - c_X^s - \sqrt{\frac{f_X}{n_X^e}}$$

[35]The result for the quadratic distance case would be:
$u_X^s(s_X, p_X^s) = -c_X^s - \sqrt{\frac{kf}{n_X^e}} - \frac{f}{16n_X^e}$

Remark: Note how this shows that the model displays indirect network effects as the reduced form expression for utility (from 'software') displays positive feedback between the *total* number of consumers on X and the utility of an *individual* on X: $\frac{d}{dn_X^e} u_X^{s'} > 0$.

Bibliography

Mark Armstrong. Competition in Two-Sided Markets. Technical report, 2005. Forthcoming in the Rand Journal of Economics (December 2006).

Mark Armstrong and Julian Wright. Two-sided Markets, Competitive Bottlenecks and Exclusive Contracts. Technical report, 2005. Forthcoming in Economic Theory.

B. Douglas Bernheim and Michael D Whinston. Exclusive Dealing. *The Journal of Political Economy*, 106(1):64–103, February 1998. ISSN 00223808.

T. Bresnahan. The Economics of The Microsoft Case, 2001. Unpublished discussion paper.

Luis Cabral. Dynamic Price Competition with Network Effects. Technical report, 2007. Unpublished.

Jay-Pil Choi. The Provision of (Two-Way) Converters in the Transition Process to a New Incompatible Technology. *Journal of Industrial Economics*, 45(2):139–153, 1997.

Jay Pil Choi. Tying in Two-Sided Markets with Multi-Homing. Working Papers 06-04, NET Institute, September 2006.

Jeffrey Church and Neil Gandal. Network Effects, Software Provision and Standardization. *Journal of Industrial Economics*, 40(1):85–103, 1992.

Jeffrey Church, Neil Gandal, and David Krause. Indirect Network Effects and Adoption Externalities, 2003. mimeo.

Stephen Davis and Kevin Murphy. A Competitive Perspective on Internet Explorer. *AER Papers and Proceedings*, 90(2):184–187, 2000.

Nicholas Economides. Symmetric Equilibrium Existence and Optimality in Differentiated Products Markets. *Journal of Economic Theory*, 47(1):178–194, 1989.

Joseph Farrell and Michael Katz. Innovation, Rent Extraction, and Integration in Systems Markets. *Journal of Industrial Economics*, 48(4):413–432, 2000.

Joseph Farrell and G. Saloner. Converters, Compatibility, and the Control of Interfaces. *Journal of Industrial Economics*, 40(1):9–35, 1992.

Franklin Fisher. The IBM and Microsoft Cases: What's the Difference. *AER Papers and Proceedings*, 90(2):180–183, 5 2000.

D. Fudenberg and J. Tirole. Pricing a Network Good to Deter Entry. *Journal of Industrial Economics*, 48(4):373–90, 2000.

R. Gilbert and Michael Katz. An Economist's Guide to U.S. v. Microsoft. *JEP*, 15(2): 25–44, 2001.

Richard J. Gilbert and Michael H. Riordan. Product Improvement and Technological Tying in a Winner-Take-All Market. *Journal of Industrial Economics*, 55(1):113–139, 03 2007.

C. Hall and R. Hall. Toward a Quantification of the Effects of Microsoft's Conduct. *AER Papers and Proceedings*, 90(2):188–191, 5 2000.

R. Jackson. Findings of Fact in the case of United States vs. Microsoft, 11 1999.

Benjamin Klein. The Microsoft Case: What Can a Dominant Firm Do to Defend Its Market Position? *JEP*, 15(2):45–62, 2001.

S. Liebowitz and S. Margolis. *Winner, Losers and Microsoft: Competition And Antitrust in High Technology*. Independent Institute, 1999.

Andrew Odlyzko and Benjamin Tilly. A refutation of Metcalf's Law and a better estimate for the value of networks and network interconnections, 2005.

Jean-Charles Rochet and Jean Tirole. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1(4):990–1029, 06 2003.

Jean-Charles Rochet and Jean Tirole. Two-Sided Markets : A Progress Report. IDEI Working Papers 275, Institut d'conomie Industrielle (IDEI), Toulouse, November 2005. Forthcoming in the RAND Journal of Economics.

M. Whinston. Tying, Foreclosure, and Exclusion. *American Economic Review*, 80(4): 837–859, 1990.

M. Whinston. Exclusivity and Tying in U.S. v. Microsoft: What We Know, and Don't Know. *JEP*, 15(2):63–80, 2001.