EUROPEAN UNIVERSITY INSTITUTE
DEPARTMENT OF ECONOMICS

EUI Working Paper **ECO** No. 2004 / 15

Propriety vs. Public Domain Licensing of Software and Research Products

ALFONSO GAMBARDELLA and BRONWYN H. HALL

BADIA FIESOLANA, SAN DOMENICO (FI)

Published in Italy in April 2004
European University Institute
Badia Fiesolana
I-50016 San Domenico (FI)
Italy

*Revisions of this work can be found at: http://emlab.berkeley.edu/users/bhhall/index.html.

Proprietary vs. Public Domain Licensing of Software and Research Products

Alfonso Gambardella

Sant'Anna School of Advanced Studies

Bronwyn H. Hall[†]

University of California at Berkeley, NBER and IFS London

Abstract

We study the production of knowledge when many researchers or inventors are involved, in a setting where there tensions can arise between individual public and private contributions. We first show, with the aid of a simple model, that without some kind of co-ordination, production of the public knowledge good (science or research software or database) is sub-optimal. Then we demonstrate that if "lead" researchers are able to establish a norm of contribution to the public good, a better outcome can be achieved. We show that the General Public License (GPL) used in the provision of open source software is one such mechanism. We then apply our results to the specific setting where the knowledge being produced is software or a database that will be used by academic researchers and possibly also by private firms, using as an example a product familiar to economists, econometric software. We conclude by discussing some of the ways in which pricing can ameliorate the problem of providing these products to academic researchers.

Keywords: open source, software, intellectual property, scientific research, databases

[†] Corresponding author:

Bronwyn H. Hall, Department of Economics, UC Berkeley, Berkeley, CA 94720-3880, USA

Fax 510 548 5561 email bhhall2@attglobal.net

Proprietary vs. Public Domain Licensing of Software and Research Products

Alfonso Gambardella and Bronwyn H. Hall

Sant'Anna School of Advanced Studies and

University of California at Berkeley, NBER and IFS

1 Introduction[1]

In the modern academic research setting, many disciplines produce software and databases as a by-product of their own activities and also use the software and data generated by others. As Dalle (2003) and Maurer (2002) have documented, many of these research products are distributed and transferred to others using institutions that range from commercial exploitation to "free" forms of open source. Many of the structures used in the latter case resemble the traditional ways in which the "Republic of Science" has ensured that research spillovers are available at low cost to all. But in some cases, moves toward closing the source code and commercial development take place, often resulting either in the disappearance of open source versions or in "forking", where an open source solution survives simultaneously with the provision of a closed commercial version of the same product. At the same time, while the production of research software or the creation of scientific databases differs in some respects from scientific research more broadly, they are areas in which the tension between the reward systems of the "Republic of Science" and the private sector are particularly obvious and important, especially when they are carried out in academic and scientific research environments.[2]

As these inputs to scientific research have become more essential and their value has grown, a number of questions and problems have arisen surrounding their provision.

First, how do we ensure that incentives are in place to encourage their supply and second, how can they be made available to researchers at a reasonable cost? How does and how should non-market production and market production of these knowledge inputs interact? In this paper, we present a simple model of the collective production of knowledge output that captures the insight that in some settings cooperative and "free" production of knowledge may break down. We then discuss how this model and its predictions might apply to the provision of scientific software and databases.

An example of the difference between free and commercial software solutions that should be familiar to most economists and scientific researchers is the scientific typesetting and word processing package ***Tex***.[4] This system and its associated set of fonts was originally the elegant invention of the Stanford computer scientist Donald Knuth, also famous as the author of the *Art of Computer Programming,* the first volume of which was published in 1969. Initially available on mainframes, and now widely distributed on UNIX and personal computer systems, ***TeX*** facilitated the creation of complex mathematical formulas in a word-processed manuscript and the subsequent production of typeset camera-ready output. It uses a set of text-based computer commands to accomplish this task rather than enabling users to enter their equations via the graphical WYSIWYG interface now familiar on the personal computer.[5]

Although straightforward in concept, the command language is complex and not easily learned, especially if the user does not use it on a regular basis. And although many academic users still write in raw***TeX*** in spite of the fact that they work on a system with a graphical interface such as Windows, there now exists a commercial program, Scientific Word, which provides a WYSIWYG environment for generating ***TeX*** documents, albeit at a considerable price when compared to the freely distributed original.

This example illustrates several features of the academic provision of software that we will develop further in our model and its discussion. First, it shows that there is willingness to pay for ease of software use even in the academic world and even if the software itself can be obtained for free. Second, the most common way in which software and databases are supplied to the academic market is a kind of hybrid between academic and commercial, where they are sold in a price-discriminatory way that preserves access for the majority of scientific users. Such products often begin as open source projects directed by a "lead" user, because the culture of open science is quite strong in the developers and participants. Nevertheless, they are eventually forced into the private sector as the market grows and non-developer users demand support, documentation, and enhancements to the ease of use.

In the following sections we first present our simple model of knowledge production when many researcher or inventors are involved, showing that without some kind of co-ordination production of the public knowledge good (science or research software or database) is sub-optimal. Then we demonstrate that if "lead" researchers are able to establish a norm of contribution to the public good, a better outcome can be achieved. We show that the General Public License (GPL) used in the provision of open source software is one such mechanism. We then apply our results to the specific setting where the knowledge being produced is software or a database that will be used by academic researchers and possibly also by private firms, using as an example a product familiar to economists, econometric software. We conclude by discussing some of the ways in which pricing can ameliorate the problem of providing these products to academic researchers.

2 A simple model of "public domain" *vs.* "proprietary" research

In this section we present a simple model of the choice of a researcher to place a given discovery or invention in the public domain as opposed to seeking private property rights on it. The model hinges on the idea that while the benefits of proprietary research stem from the rents that can be enjoyed from the sales or the license of the invention, the benefits from contributing to the public good stem from various sources. Apart from a system of values that prizes contributions to public domain knowledge, research outcomes that fall into the public domain produce visibility (which potentially increases future incomes) or non-pecuniary benefits like fame and glory. We also like to think that our model embraces a wide set of situations. It can be thought of as a model that mimics the choice of academic scientists to publish their research findings vis-à-vis holding patents or other property rights on them (Dasgupta and David, 1994), or software developers who contribute to open source software as opposed to patenting their programs (Lerner and Tirole, 2002), or user-inventors who may transfer their inventions to the producers or market them using intellectual property protection (Von Hippel, 1988; Von Hippel and Von Krogh, 2003; Harhoff, Henkel, and Von Hippel, 2003).

Although quite simple, the model produces some interesting insights.[6] First, it highlights the inherent co-ordination failure of this problem. Most notably, if the potential contributors to the public good only look at their individual contribution, they may well find it profitable to deviate by privatizing their output. This is because the individual contribution to the public good is small. Hence, as one individual deviates, and she holds the behaviour of the others as given, she will lose only very little from her reduced contribution. By contrast, she can enjoy a discrete increase in income from selling her invention onto the market place, possibly at some non-competitive price as implied by the fact that she may hold property rights on it. By the same token, moving from property rights to public domain implies only a small gain in terms of a better public good, while losing the rents from the privatization of research. The co-ordination failure arises because collectively the individuals produce sizable public domain knowledge. Hence, if they could co-ordinate and stick to the production of the public good, they would in the end be better off. In our model we obtain that with no co-ordination fewer people end up putting their findings in the public domain than if they co-ordinate.

This is the classical result by Mancur Olson (1971) that unless there is co-ordination among the independent agents, the collective action is hard to sustain. In our terms, this suggests that there is an asymmetry between public domain and private knowledge production. The latter is easier to sustain than the former. Therefore the role for policy is much more manifest when it is desirable to enhance public domain rather than privately owned knowledge, a point made long ago by Nelson (1959). This also suggests why there is a tendency for certain types of knowledge to move out of the public domain over time (e.g. the shift of academic scientists to privatization when knowledge becomes closer to economic applications, as for example in biotechnology or software) rather than the other way around. Moreover, this suggests that non-economic factors – be them the norms of open science (Dasgupta and David, 1994) or the principles of open source software (Raymond, 1999), or other mechanisms – are crucial to sustain the production of knowledge under public domain.[7]

Our model first shows that in equilibrium a share of potential researchers work under public domain rules and the remaining researchers work under proprietary rules. For example, this mimics cases in which some software programs are sold under open source, while others are sold under traditional commercial rules, as we shall see with some examples in later sections. Second, we characterize the factors that can induce higher or lower share of researchers working under public domain in equilibrium. A fairly natural one is that if the potential economic rents that can be gained from the privatization of knowledge increase, the share of researchers working under public domain decreases. Thus, for example, stronger Intellectual Property Rights (IPRs), larger markets (with implied higher rents) for a given invention, closer distance between the contributions of the researchers to profitable market opportunities (e.g. in academic biotechnology today, as well as in some other sciences), are all factors that can put serious pressures on public domain research. To put it in a different way, one needs tighter non-economic co-ordination devices (e.g. a stronger reliance on a system of norms and values) to keep the public domain equilibrium viable.

In our model, the key mechanism that sustain a public domain equilibrium with a higher number of researchers is the co-ordination of a sizable number of individuals who stick to the public domain diffusion of their results. The question is then what can give rise to this co-ordination. We develop a simple extension of the model that shows that the principle of the Generalized Public License (GPL) in open source software can provide such a co-ordination. As discussed for instance by Lerner and Tirole (2002), the GPL is a license whereby the producer of an open source program requires that all modifications and improvements of the program be subject to the same rules of openness, most notably the source code of all the modifications ought to be made publicly available like the original program. We show that this creates the necessary co-ordination to solve the Mancur Olson problem and sustain the public domain equilibrium. As we shall discuss, we argue that this offers a new perspective about how to support public domain knowledge, viz. you sometimes need to enhance public domain knowledge via some institutional mechanism. The GPL is one such mechanism.

Finally, our model abstracts from many specificities of the public domain *vs.* proprietary research approach, which may indeed yield some important insights. In this paper, we deliberately focused on some simpler features of this choice, as we wanted to highlight a few general aspects of the problem that we believe to be relevant in this context and that we think had not yet been emphasized enough by the literature.

The basic model

Assume that a set of individuals work on a set of research projects. Each individual undertakes some projects under proprietary research (PR) rules and others under public domain (PD) rules. In the latter case, she places her findings in the public domain, and enjoys no economic rents. In the former case, she does not make her findings public, she seeks property rights on them, and she enjoys profits. We assume that the individuals benefit from the pool of public domain knowledge. There can be many reasons for this. They could have values that prize the growth of public knowledge (they "consume" public knowledge); or they benefit from public knowledge domain as an input for their own activities. Clearly, the body of public knowledge is larger the higher the number of researchers that, in any given field, choose to make their findings public instead of keeping them private. Ultimately, the (indirect) utility function of the individual is

$$U = \sum_{i \in PR} \left(X_i + \pi_i \right) + \sum_{i \in PD} X_i \qquad (1)$$

where *i* index the research projects, PD and PR represent the set of projects carried out under public domain and proprietary research, X denotes the benefits from the available public domain knowledge in any given field, and π is the profits earned from her PR projects. We also assume that there is heterogeneity across individuals with $\pi_i \sim F_i(\pi)$, π_i bounded. The profits π_i could well be negative if we allow for the possibility that there are costs to the researcher of transferring the knowledge to private use. For instance, they may need to set up a new company, and this may involve spending time away from other activities. Thursby, Jensen, and Thursby (2001) report that faculty frequently fail to disclose inventions to their university's technology transfer office because they fear having to spend time on the commercialization activity. Note also that the distribution of the profits across individuals is indexed by *i*, which means that the distribution can differ across projects.

Each individual will carry out under PR all the projects such that $X(n) \geq X(n-1) + \pi$, where for simplicity we suppressed the index *i* for projects, and where $n-1$ is the number of "other" researchers that have chosen to work under PD rules in that particular field. The share of researchers in the field that work under PD is then $F(\Delta X_n)$, where $\Delta X_n \equiv X(n) - X(n\text{-}1)$. We also make some innocuous assumptions:

$$\frac{\partial X}{\partial n} \geq 0 \qquad (2a)$$

$$\frac{\partial \Delta X_n}{\partial n} \leq 0; \quad \Delta X_n \to 0 \text{ as } n \to \infty \qquad (2b)$$

Assumption 2a) is the one stated earlier. It says that the benefits from the public domain knowledge do not decrease as more people work under PD rather than PR. Assumption 2b) states that the contributions to public domain research exhibits diminishing returns. The higher the number of individuals working under PD in the field the smaller the contribution of any additional individual working under PD in that field. Moreover, the individual contributions to the public good become negligible as *n* becomes large.

Case 1: No Co-ordination

We first look at the case in which there is no co-ordination among the individuals. The equilibrium share of PD researchers in a given field is determined by $F(\Delta X_n) = n/N$, where *N* is the total number of researchers. Figure 1 depicts this equilibrium. Point *E* in Figure 1 is an equilibrium because if the number of researchers working under PD increase beyond the equilibrium level n^e, the share of researchers with $\pi \leq F(\Delta X_n)$ decreases, which means that there are fewer individuals that prefer to work under PD than in equilibrium. But this is a contradiction, and therefore there is no incentive to deviate. The reasoning is analogous if we consider deviations from PD to PR in equilibrium.

[Figure 1 about here]

It is also easy to see that the share of researchers working under PD decreases if the economic profitability of the research in the field increases. This can be thought of as a first-order stochastic increase of the distribution of profits $F(\cdot)$. That is, π is distributed according to a function such that for any given π_o the probability of $\pi \leq \pi_o$ is smaller. For example, this can happen if there is a common shock across the individuals that makes research in a certain field closer to commercial applications. It could also be the result of the availability of stronger intellectual property rights or of a change in university policy towards the use of IPRs, or of changes in particular research areas such as the life sciences or software, in which academic research has become closer to potential commercial applications.

In Figure 1 this implies a downward shift of $F(\Delta X_n)$ for any given ΔX_n. It is easy to see that this implies that the equilibrium number of PD researchers decreases. As a matter of fact, there is fairly widespread evidence that in fields like software or biotechnology there is growing pressure on academic researchers to place their findings in a proprietary regime. Also, our examples in the later sections show that shifts from academic to commercial software are more prominent when the market demand for the products increase, which raises the profitability of the programming efforts. Finally, there are several accounts of the fact that tension between industrial research and academic norms become higher if university access to IPRs is increased (Cohen, Florida, Randazzese, and Walsh 1998; Hall, Link, and Scott 2001; Hertzfeld, Link, and Vonortas 2004; Cohen, Florida, and Randazze 2004). As these authors report, such tensions have already begun in the US, as the latter country has pioneered the trend towards stronger IPRs and the use of intellectual property protection by universities, but they are becoming more pronounced in Europe as well, as European universities follow the path opened up by the US system (Geuna and Nesta 2004). Collins and Wakoh (1999) describe similar changes in Japan and describe how the regime shift to patenting by universities is inconsistent with the previous system of collaborative research with industry in that country, implying increasing stress for the system.

By contrast, the share of researchers working under PD increases in highly productive scientific fields, such as new or more fertile areas of research where individual academic contributions are more important. Suppose for example that in a given field the increase in PD output from the *n*th researcher's contribution (ΔX_n) increases for any level of *n*. In Figure 1 this would shift the $F(\cdot)$ curve upward, with an implied increase in the equilibrium *n*.

Case 2: Co-ordination

Our simple set-up is suggestive of why co-ordination can help raise the share of PD researchers. The argument is straightforward. Suppose that each researcher can co-ordinate a total of *v* individuals rather than just himself. By this we mean that he knows that his decision to switch to PR or PD implies that $v - 1$ other individuals also switch.[9] All the researchers with $\pi \leq \Delta X_n^v$ will then choose to work under *PD*, where $\Delta X_n^v \equiv X(n) - X(n\text{-}v)$. But assumptions 2a) and 2b) imply that $\Delta X_n^v > \Delta X_n$. As a result, in Figure 1 the $F(\cdot)$ curve shifts upward, and the equilibrium *n* increases.

The Generalized Public License (Copyleft) as a Co-ordination Device

But how does this co-ordination take place? We argue in this section that the copyleft system (Generalized Public License, GPL) is one way to obtain it. As noted earlier, and as discussed by Lerner and Tirole (2002), the GPL was first implemented by the MIT software programmer Richard Stallman, who devised a license for his software program that imposed that all source codes based on modifications of his initial program be made freely available under the same conditions as his initial source code. To model the implications of the GPL, suppose that one researcher can potentially launch a new project. He has three decisions to take: a) whether to launch the project or not; b) if he launches it, whether to do it under PD or PR; c) if PD, whether to attach a GPL to it.[10] The other researchers then have to decide whether to join this project or not. We use the same notation as in the previous section. Thus, $X(n)$ is the public benefit accruing to the community of researchers who join this project if n of them operate under PD, and π is the private profit that they enjoy if they choose to operate under PR. We also assume that all the researchers who undertake this project, including the one who launches it, have an opportunity cost. We label it as B, and assume that B is distributed across the individuals as $B \sim G(B)$, B bounded.[11]

To solve this model, we work backwards to the decision of the researcher to launch the new project. We start with the case in which he has launched the project, works under PD and attaches a GPL to the project. The latter implies that anyone who would like to contribute to this project cannot privatize his efforts. As a result, each researcher will decide whether to join or not according to whether his opportunity cost $B \leq X(n)$. As in the earlier case (Figure 1), the equilibrium n is defined by $G(X(n)) = n/N$. Here however, $G(X(n))$ increases rather than decreasing with n. This means that: a) there can be multiple equilibria; and b) the equilibrium n is the one where $G(X(n))$ cuts the n/N line from above (that is, $\frac{\partial G}{\partial n} \leq \frac{1}{N}$). Figure 2 depicts the possible equilibria in this case. We label the equilibrium n under GPL as n^G.[12]

[Figure 2 about here]

Suppose that the leader did not attach a GPL to the project. In this case, the potential contributors no longer check only $B \leq X(n)$. They will work on the project under PD if *[B≤X(n)* *__and__* $\pi \leq X(n) - X(n\text{–}1)]$. They must now prefer PD to PR as well as satisfying the opportunity cost condition. Not using a GPL adds a new constraint on the choice set of the individual, as in the previous section. As a result, the set of individuals who works on the new project under PD can only be smaller without a GPL. Define $\Gamma(B,\pi)$ to be the joint probability distribution of *B* and π.[13] The equilibrium *n* in this case, which we label n^{NG} ("no GPL"), is the one that solves $\Gamma(X_n, \Delta X_n) \equiv \Gamma(n) = n/N$, with $\frac{\partial \Gamma}{\partial n} \leq \frac{1}{N}$. But $\Gamma(n) \leq G(n)$ for any *n*. As a result, for any equilibrium in Figure 2, it must be that $n^{NG} \leq n^{G}$; that is, the GPL implies a larger number of researchers working under PD at any level of *n* and therefore in equilibrium.

We now ask whether the initial researcher will launch the new project under GPL. If he chooses to operate under PD, the answer is clearly yes. His utility under PD and GPL is $X(n^G)$. Since $n^G \geq n^{NG}$, this cannot be smaller than $X(n^{NG})$. Of course, if he chooses to operate under PR, GPL would be his best choice as well. This is because $X(n^G - 1) + \pi \geq X(n^{NG} - 1) + \pi$. It is optimal to privatize while others operate under public domain. We rule this possibility out by assumption, because it is hard to sustain an action whereby the researcher chooses to privatize his own results while imposing that others make their findings public. We therefore assume that if people work under *PR* they cannot attach a GPL to subsequent additions to the stock of public knowledge.

Working backward, the previous stage is the researcher's choice whether to carry out the project under PD (and GPL) or PR (and no GPL), conditional of having chosen to do the project. He will choose PD and GPL if $X(n^G) \geq X(n^{NG} - 1) + \pi$. Note that compared to the no co-ordination case discussed in the earlier section, the distance between the arguments of *X* under the two regimes is no longer one unit (i.e. the researcher himself), but a set of researchers $n^G - n^{NG}$ in addition to the researcher himself.

This distance resembles the v discussed in the section where we allow for the possibility of co-ordination. Will the researcher launch the project at all? Yes, if $B \leq X(n^G)$ and $B \leq X(n^{NG} - 1) + \pi$.

We are now ready to characterize the conditions under which a researcher with a potential new project will: i) launch it under PD and GPL; ii) launch it under PR (with no GPL); iii) not launch it. These are:

$$\text{(PD and GPL)} \qquad B \leq X(n^G) \ \underline{and} \ \pi \leq X(n^G) - X(n^{NG} - 1) \qquad \text{(3a)}$$

$$\text{(PR and no GPL)} \qquad B \leq X(n^{NG} - 1) + \pi \ \underline{and} \ \pi \geq X(n^G) - X(n^{NG} - 1) \qquad \text{(3b)}$$

$$\text{(No project)} \qquad B \geq X(n^G) \ \underline{and} \ B \geq X(n^{NG} - 1) + \pi \qquad \text{(3c)}$$

Now suppose that there was no GPL. It is not difficult to see that the conditions that would characterize the launch of the project will be the same as (3a)-(3c), with n^{NG} in lieu of n^G wherever the latter appears. Given that $n^G \geq n^{NG}$, the result is that the possibility of a GPL implies the following: a) there will be a greater number of new projects in equilibrium; b) the new projects are more likely to be launched under PD; and c) if $X(n^G)$ is large relative to $X(n^{NG})$, as is likely, the projects added are disproportionately those that were not privately profitable. The intuition about these results is straightforward given our discussion so far. The GPL provides the necessary commitment to allow for a greater co-ordination of researchers who commit to contribute to a larger public good. As a result, the value of the new projects increase (hence more of them are carried out), and the value of projects under PD increases (hence more of them are carried out).

These results rely on the fact that there is enforcement of the GPL. Since the GPL is not like a patent or a copyright, which are enforced by law, one may question whether the copyleft system can actually be enforced. However, in some settings people seem to abide by the copyleft rules, as Lerner and Tirole (2002) have noted, in spite of the lack of legal enforcement. In many situations, there may be a reputation effect involved when the copyleft agreement is not complied with. But if this is true, why is it then needed in the first place? Without an explicit copyleft license it may not be clear to the additional contributors whether the intention of the initial developers of the project was to keep it under PD or not. But if their will is made explicit, deviations may be seen as an obvious and explicit challenge to the social norms, and this may be sanctioned severely by the community.

A related point is that the literature has typically been concerned with the need to protect the private property of knowledge when this is necessary to enhance the incentives to innovate. The inherent assumption is that when it is not privately protected, the knowledge is by default public, and it enriches the public domain. Yet, our model points out that this is hardly true. The public nature of knowledge needs itself to be protected when commitments to the production of knowledge in the public domain is socially desirable. In other words, there is a need for making it explicit that the knowledge has to remain public, and this calls for positive actions and institutions to protect it. Not allowing for private property rights on some body of knowledge is not equivalent to assuming that the knowledge will be in the public domain. One may then need to assign property rights not just to private agents, but also to the public. For example, the IPRs are typically thought as being property rights to private agents. But we also need to have institutions that preserve the public character of knowledge. The copyleft license is a beautiful example of this institutional device. A natural policy suggestion is therefore to make it legal and enforceable as copyright, patents and other private-based IPRs.

Finally, our model is suggestive of when the copyleft license makes a real difference. Suppose that B and π are positively correlated. This means that individuals with high B tend to have high π. Recall that a high B may arise because of a highly valuable public good project (high X) or a high profit in another field. Thus, for example, a high positive correlation between the two could arise because there is some common element across projects that makes the individual effective in commercializing knowledge in any field – e.g. the researcher is linked to start-up companies that can commercialize any kind of project; or his institution (university or else) encourages the commercialization of knowledge (has a good licensing office); or he is not that keen about keeping science public. In this case the copyleft agreement will not make a big difference. The intuition is that the individuals who will join the new project are those with low B. But they also have low π, and hence all those who join are likely to do it under PD rather than PR. Hence, there is no need for an explicit license to "force" them to work under PD. In terms of our model, a high correlation between B and π means that $\Gamma(B,\pi)$ is close to $G(B)$, or that the event $[B \leq X(n); \pi \leq X(n)–X(n–1)]$ is almost as likely as the event $[B \leq X(n)]$ because a great share of the individuals with $B \leq X(n)$ (low B) also exhibit $\pi \leq X(n) – X(n–1)$ (low π). As a consequence, $\Gamma(n)=n/N$ and $G(n)=n/N$ yield equilibrium n that are close to each other. The distance $n^G – n^{NG}$ is not large.

Suppose instead that B and π are not correlated, or that they are negatively correlated. This means that individuals with low B may have high π. As a result, individuals who benefit from participating in the project may prefer to participate under PR rather than PD. Copyleft can then make a difference as it "forces" the people who find it profitable to participate to do so under PD rather than PR. In short, $n^G – n^{NG}$ is likely to be high.[14]

One way to think of independence between B and π, as opposed to positive correlation, is that in the latter case the new projects are not very novel. The factors that affect the individual benefits in their current activity are similar to those that affect their potential profit in the new project. By contrast, when the projects are radically new, the opportunities of the individuals change substantially, and the researchers who might profit the most from the new projects can be different from those who benefited the most in the old projects. For example, new skills, or new forms of learning are necessary in the new fields, and the people who have made substantial investments in the old projects may have greater difficulties in the new areas. (See, for example, Levinthal and March, 1993.) In these cases, researchers with low B may instead find that they have great opportunities to commercialize knowledge in the new fields (high π). Thus, copyleft agreements are more likely to be useful when the project is radically new rather than incrementally different from previous projects, and when it is socially desirable to run these projects under PD.

3 Complementary investments in open source production

An important feature of traditional open source or academic software production that we alluded to in the introduction is that it normally requires additional investments that enhance the usefulness and value of the scattered individual contributions, or it simply requires investments to combine them. For example, while several individuals can contribute to the development of a whole body of scientific knowledge, there must be some stage in which the "pieces" are combined into useful products, systems, or transferable knowledge. Some scientists or most likely some specialized agents, i.e. academic licensing offices or firms, normally perform this function. A typical example is when scientific knowledge needs substantial downstream investments to become economically useful technologies or commercializable products. Thursby, Jensen, and Thursby (2001) report that such is often the case for university research outputs. The latter activities are normally performed by firms. In software additional investments are often required to enhance the usability of the software for those who did not develop it, and to produce documentation and support. The need for additional investments in open source production, or more generally in tasks that rely on public domain knowledge, has some specific implications that we want to discuss in this section. In our companion paper (Gambardella and Hall, 2004) we develop a model that addresses this point. Here we simply discuss some key aspects arising in part from that model.

The problem is that the (downstream) "assembling" agent needs some profits in order to carry out the investments that are necessary to produce the complementary downstream assets of the good. Since the downstream assembling agents are typically firms, we now refer to them as the latter. There are two questions. First, the firm needs to obtain some economic returns to finance its investment. Clearly, there are many ways to moderate its potential monopoly power so that the magnitude of the rents will be sufficient to make the necessary investments but not high enough to produce serious extra-normal profits. However, it would be difficult for the firm to obtain such rents if it operated under perfect competition, or if it operated under an open, public domain system itself.

The second question is more subtle. As modelled in our companion paper, the firm uses the PD contributions of the individual agents (software programmers, scientists, etc.) as inputs in its production process. If these contributions are freely available in the public domain, and particularly they are not available on an exclusive basis, many downstream firms can make use of them. As a result, the downstream production can easily become a free entry, perfectly competitive world, with many firms having access to the widely available knowledge inputs. If so, each firm could not make enough rents to carry out the complementary investments. This would be even harder for the individual knowledge producers who are normally scattered and have no resources to cover the fixed set-up costs for the downstream investments. The final implication is that the downstream investments will not be undertaken, or they will be insufficient. Of course, there can be other factors that would provide the firms with barriers to entry, thereby ensuring that they can enjoy some rents to make their investments. However, in productions where the knowledge inputs are crucial (e.g. software), the inability to use them somewhat exclusively can generate enough threats of widespread entry and excessive competition to discourage the complementary investments.

Paradoxically, if the knowledge inputs were produced under proprietary rules, the producers of them could charge monopoly prices (e.g. because they could obtain an exclusive license). This raises the costs of the inputs. In turn, this raises the barriers to entry in the downstream sector, and adjust the level of downstream investment upward. In other words, if the inputs are freely available there could be excessive downstream competition, which may limit the complementary investments. If they are offered under proprietary rules, the costs of acquiring the inputs are higher, which curbs entry and competition, and allows the downstream firms to make enough rents to carry out such investments.[15]

But the privatization of the upstream inputs has several limitations. For one, as Heller and Eisenberg (1998) have noted, the complementarity among the "pieces" of upstream knowledge produced by the different individuals can give rise to the so-called problem of the anti-commons. That is, after all the other rights have been collected under a unique proprietorship, the final owner of a set of complementary inputs can enjoy enormous monopoly power. This is because by withholding his own contribution, he can forestall the realization of the whole technology, especially when the complementarity is so tight that each individual contribution is crucial to make the whole system work. The possibility of ex-post hold-up can discourage the effort to collect all the complementary rights ex-ante, and therefore prevent the development of the technology. Another limitation of the privatization of the upstream inputs is the one discussed in the previous section. With copyleft agreements, more people can contribute to the public good, which enhances its quality. The decentralized nature of the process by which scientists or open source software producers operate has typically implied that, with public domain knowledge production, the network of contributors to a given field can be so large that the overall improvements can be higher than what can be obtained within individual organizations, including quite large ones. Some evidence that open source projects also increase the quality of software output has been supplied by Kuan (2002).

One solution to the problem of paying for complementary downstream investment is allowing for property rights, and particularly intellectual property rights, on the innovations of the downstream producer. This would of course raise its monopoly power and therefore curb excessive competition. At the same time, it avoids attaching IPRs to pieces of upstream knowledge thereby giving rise to the problems of the anti-commons, or to reduced quality of the upstream knowledge. In addition, the downstream producer would enjoy rights on features of the innovation that are closer to his own real contribution to the project, that is the development of specific downstream investments. Clearly, this also implies that the IPRs thus offered are likely to be more narrow, as they apply to downstream innovations as opposed to potentially general pieces of knowledge upstream. At the same time, they are not likely to be as narrow as in the case of small individual contributions to an open software module or a minor contribution to a scientific field, which can give rise to the fragmentation and hold-up problems discussed earlier.

4 Academic software and databases

In this section we draw some implications for the provision of scientific software and databases from the model and discussion in the previous two sections and then go on to discuss the possible modes in which they could be provided. First, this type of activity is more likely to be privatized than scientific research itself because there is greater and more focused market demand for the product, because norms are weaker due to weaker reputation effects, and because there are more potential users who are not inventors (and do not participate in the production of the good). Second, there could easily be both public and private provision at the same time, because such an equilibrium can be sustained when there are different communities of researchers with different norms. Third, as the market for a particular product grows, privatization is likely simply because the individual's discrete return to privatization has increased. Finally, when the components to a valuable good are produced under PD, free entry in the downstream industry producing a final good based on those components implies too few profits for those undertaking investments that will enhance the value of the good. The final producers have to earn some rents to be able to make improvements beyond the mere availability of research inputs.

The production of research software and databases is different from that of scientific research more broadly, in that it is a by-product of the central activity undertaken by researchers. This fact has implications for its production in a patronage or public-funding environment that follows the norms of open science:

1) Usually the production of software and databases is not salient for the funding body, in the sense that there will be a tendency not to fund development fully and not to fund the necessary maintenance to make it useful for others subsequent to its first use. Granting agencies often have a preference for new initiatives, which disfavours ongoing development.
2) The incentive system of open science, which is based on reputation and priority for scientific discoveries, is ill-suited to support the development of software and databases. It is particularly poor at generating maintenance activities and features

that make the data or software useful for others such as user interfaces and documentation, because these support activities do not create the kind of discoveries that are rewarded effectively in these regimes.

3) The spillover benefits of software and database development are largely one-way and therefore not usually as subject to reciprocal exchange as the generation of research results.

4) Supplying this type of information product has traditionally required direct transfer activities (the mailing of tapes or CDs, sending of files over the internet, etc.), unlike the output of research itself (largely available in academic libraries), so the users are identifiable (and therefore can be charged, often in a price-discriminating manner).

The privatization of scientific databases and software has both advantages and disadvantages. With respect to the latter, David (2002) has emphasized the negative consequences of the privatization of scientific and technical data and information. One of the most important drawbacks is the increase in cost, sometimes substantial, to other scientists, researchers or software developers for use of the data in ways that might considerably enhance public domain knowledge. A second is that the value of such databases for scientific research is frequently enhanced by combining them and/or using them in their entirety for large scale statistical analysis, both of which activities are frequently limited when they are commercially provided.[16] Maurer (2002) gives a number of examples of privatized databases that have somewhat restricted access for academic researchers via their pricing structure or limitations on reuse of the data, such as Swiss-PROT, Space Imaging Corporation, Incyte, and Celera. In this issue, David (2004) cites the case of the privatization of Landsat images under the Reagan administration, which led to a tenfold increase in the price of an image. In terms of our model, the potential to privatize scientific and technical data and information implies that a smaller number of researchers *n* will contribute to the public good, with implied smaller *X(n)* than if such an opportunity was not available.

At the same time, a common argument in favour of privatization of databases is that it helps in the development of a database producing industry, and more generally of an industry that employs these data as inputs. A similar argument can be used more broadly for software. For example, the recent European Directive that defines the terms for the patenting of software in Europe (European Commission, 2002) was largely justified by the argument that it would encourage the formation of a software industry in niches and specialized fields. Although it is sometimes true that exclusivity can have positive effects on the provision of information products, it is also true that there can be drawbacks like those suggested earlier (fragmentation of IPRs, little contribution to public domain knowledge, restricted access when welfare would be enhanced with unlimited access) to the privatization of knowledge inputs. At times, one can obtain similar advantages by allowing for the privatization of the outputs that can be generated using the database or software in question. That is, discovery of a useful application associated with a particular gene that is obtained by use of a genomic database is patentable in most countries. Or, in the case of the econometric software example used later in the paper, consulting firms such as Data Resources, Inc. or Chase Econometrics marketed the results of estimating econometric models using software whose origins were in the public domain. Following our earlier argument, by allowing for the privatization of the downstream output we make it possible for the industry to obtain enough rents to make the necessary complementary investments, while avoiding the limitations of privatizations in the upstream knowledge.

There are, however, limits to this particular strategy for ensuring that scientific databases and software remain in the public domain at the same time that downstream industries based on these freely available discoveries can earn enough profit to cover their necessary investments. The difficulty of course is that in the case of generally useful information products, a firm selling a particular product one of whose inputs is an upstream academic product has no reason to undertake the enhancements to the upstream product that would make it useful to others, unless the firm can sell the enhanced product in the marketplace. But this is what we were trying to avoid, and what is ruled out by a GPL. We now turn to a discussion of an alternative way in which such goods can be provided.

The production of information products including software and databases has always been characterized by large fixed costs relative to marginal cost, but the cost disparity has grown since the advent of the internet. In practice, the only real marginal costs of distribution arise from two sources: the support offered to individual users (which in many cases has been converted into a fixed cost by requiring users to browse knowledge bases on the web) and the congestion costs that can occur on web servers if demand is too great.[17] Standard economic theory tells us that when the production function for a good is characterized by high fixed costs and low marginal costs, higher welfare can often be achieved by using discriminatory pricing, charging those with high willingness to pay more in order to offer the good to others at lower prices, thus increasing the overall quantity supplied. The problem with applying this mechanism generally is the difficulty of segmenting the markets successfully and of preventing resale.

In the case of academic software and databases, however, it is quite common for successful price-discriminating strategies to be pursued.[18] There are several reasons for this: 1) segmentation is fairly easy because academics can be identified via addresses and institutional web information; 2) resale is difficult in the case of an information product that requires signing on to use it and also probably not very profitable; 3) the two markets (academic and commercial) have rather different tastes and attitudes toward technical support (especially towards the speed with which it is provided) so the necessary price discrimination is partly cost-based.

5 Case Study: Econometric Software Packages

As an illustration of the pattern of software development in the academic arena, we present some evidence about a type of product familiar to economists that has largely been developed in a university research environment but is now widely available from commercial firms: packaged econometric software. Our data are drawn primarily from the excellent surveys on the topic by Charles Renfro (2003a, b). We have supplemented it in places from the personal experience of one of the co-authors, who participated in the activity almost from its inception. The evidence supplied here can be considered illustrative rather than a formal statistical test of our model, since the sample is relatively small. To form a complete picture of the phenomenon of software and database commercialization in academia, it would be necessary to augment our study with other case studies. For example, see Maurer (2002) for a good review of methods of database provision in scientific research.

Econometric software is very much a by-product of the empirical economic research activity, which is conducted largely at universities and non-profit research institutions and to a lesser extent in the research departments of banks and brokerage houses. It is an essential tool for the implementation of statistical methods developed by econometric theorists, at least if these methods are to be used by more than a very few specialists. To a great extent, this type of software originated during the 1960s, when economists began to use computers rather than calculating machines for estimation, and for the first time had access to more data than could comfortably be manipulated by hand. The typical such package is implemented using a simple command language and enables the use of a variety of modelling, estimating and forecasting methods on datasets of varying magnitudes. Most of these packages are now available for use on personal computers, although their origins are often a mainframe computer implementation. For a complete history of the development of this software, see Renfro (2003b).

Like most software, econometric software can be protected via various IP measures. The most important is a combination of copyright (for the specific implementation in source code of the methods provided) and trade secrecy (whereby only the "object" code, or machine-readable version of the code, is released to the public). This combination of IP protection has always been available but has only become widely used during the personal computer era. Prior to that time, distributors of academic software usually provided some form of copyrighted source code for local installation on mainframes, and relied on the fact that acquisition and maintenance were performed by institutions rather than a single individual to protect the code. This meant the source code could be modified for local use, but because the size of the potential market for "bootleg" copies of the source was rather small, piracy posed no serious competitive threat. The advent of the personal computer, which meant that in many cases software was being supplied to individuals rather than institutions changed this situation, and today the copyright-trade secrecy model is paramount.[19] Thus it is possible to argue that developments in computing have made the available IP protection in the academic software sector stronger at the same time that the potential market size grew, which our model implies will lead to more defection from public domain to proprietary rules.

In Table 1, we show some statistics for the 30 packages identified by Renfro. The majority (20 of the 30) have their origins in academic research, either supported by grants or, in many cases, written as a by-product of thesis research on a student's own time.[20] A further 5 were written specifically to support the modelling or research activities of a quasi-governmental organization such as a central bank. Only 5 were written with a specific commercial purpose in mind. Two of those 5 were forks of public domain programs, and in contrast to those of academic origin (whose earliest date of introduction was 1964 and whose average date was 1979), the earliest of the commercial programs was developed in 1981/82, a date that clearly coincides with the introduction of the non-hobbyist Personal Computer. Notwithstanding the academic research origin of most of these packages, today no less than 25 out of the 30 have been commercialized, with an average commercialization lag of 9 years.

Reading the histories of these packages supplied in Renfro (2003b), it becomes clear that although many of them had more than one contributor, normally there was a "lead user" who coordinated development, the identity of the "lead user" occasionally changing as time passed. Most of the packages had their origins in the solution of a specific research problem (e.g., the development of LIMDEP for estimation of the Nerlove and Press logit model, or the implementation of Hendry's model development methodology in PCGive), but were developed, often through the efforts of others besides the initial inventor, into more general tools.

These facts clearly reflect the development both of computing technology and of the market for these kinds of packages. As predicted by our model, growth in the market due to the availability of personal computers and the growth of the economics profession as whole has caused the early largely open source development model of the 1960s to become privatized. Nevertheless, there remain five programs that are supplied for free over the internet; of these three had their origins prior to 1980 and the other two are very recent. As our model suggests, not all of the individuals in the community shift to the private system, and the equilibrium *n* can well be between zero and *N*. Interestingly, only one of the five is explicitly provided with a GPL attached. A quote from one of the author's websites summarizes the motivation of those who make these programs available quite well:

> **Why is EasyReg free?**
>
> EasyReg was originally designed to promote my own research. I came to realize that getting my research published in econometric journals is not enough to get it used. But writing a program that only does the Bierens' stuff would not reach the new generation of economists and econometricians. Therefore, the program should contain more than only my econometric techniques.
>
> When I taught econometrics at Southern Methodist University in Dallas in the period 1991-1996, I needed software that my graduate students could use for their exercises. The existing commercial software was not advanced enough, or too expensive, or both. Therefore, I added the econometric techniques that I taught in class first to SimplReg, and later on to EasyReg after I had bought Visual Basic 3.
>
> Meanwhile, working on EasyReg became a hobby: my favourite pastime during rainy weekends.
>
> When I moved to Penn State University, and made EasyReg downloadable from the web, people from all over the world, from developing countries in Asia and Africa as well as from western Europe and the USA, wrote me e-mails with econometric questions, suggestions for additions, or just saying "thank you". It appears that a lot of students and researchers have no access, or cannot afford access, to commercial econometrics software. By making EasyReg commercial I would therefore let these people down.
>
> There are also less altruistic reasons for keeping EasyReg free:

> * By keeping EasyReg free my own econometric work incorporated in EasyReg will get the widest distribution.
>
> * I will never be able to make enough money with a commercial version of EasyReg to be compensated for the time I have invested in it.
>
> * Going commercial would leave me no time for my own research.[21]

Indeed, the second statement suggests that one reason to leave the software in the public domain was that the researcher's commercial profits (π in our model) were not large enough. Likewise, the third statement suggests that the researcher cared about research and this was an important reason for not privatizing it. In our model π is the relative utility of commercial profits vis-à-vis the preference for research. Hence, the third statement is also suggestive of a low π of the individual. In sum, the model's prediction that both private and public modes of provision can co-exist when at least some individuals adhere to community norms is borne out, at least for one example.

We also discussed explicitly the role of complementary services or enhanced features for non-inventor users in the provision of software. This is clearly one of the motivations behind commercialization, as was illustrated by the example of ***TeX***. Table 2, which is drawn from data in Renfro (2003a) attempts to give an impression of the differences between commercialized and non-commercialized software, admittedly using a rather small sample. To the extent that ease of use can be characterized by the full WIMP interface, there is no difference in the average performance of the two types of software. The main differences seems to be that the commercialized packages are larger and allow both more varied and more complex methods of interaction. Note especially the provision of a macro facility to run previously prepared programs, which occurs in 84 per cent of the commercial programs, but only in 2 out of the 5 free programs. Such programs are likely to require more user support and documentation, because of their complexity, which increases the cost of remaining in the *PD* system. In short, as our earlier discussion and our model in Gambardella and Hall (2004) suggested, a commercial operation, which is likely to imply higher profits, also provides a greater degree of additional investments beyond the mere availability of the research inputs.

To summarize, the basic predictions of our model, which are that participants in an open science community will defect to the private (IP-using) sector when profit opportunities arise (e.g. the final demand for the product grows, or IP protection becomes available), are confirmed by this example. We also find some support for the fact that commercial operations are likely to undertake more complementary investments than pure open source operations. We do not find widespread use of the GPL idea in this particular niche market yet, although use of such a license could evolve. In the broader academic market, Maurer (2002) reports that a great variety of open source software licenses are in use, both viral (GPL, LPL) and non-viral (BSD, Apache-CMU).

Pricing

Our model does not explicitly incorporate all the factors that are clearly important in the case of software and databases. Specifically, one area seems worthy of further development. We did not really model the competitive behaviour of the downstream firms in the database and software industries. In practice, in some cases, there is competition to supply these goods, and in others, it is much more common for the good to be supplied at prices set by a partially price-discriminating monopolist. We report the evidence on price-discrimination for our sample briefly here.

Table 3 presents some very limited data for our sample of 30 econometric software packages. Of the 30, 5 are distributed freely and a further 8 are distributed as services, possibly bundled with consulting (such sales are essentially all commercial); this is the “added value” business model discussed earlier. Of the remaining 17, we were able to collect data from their websites for 15. Of these only 2 did not price discriminate, 3 discriminate by the size and complexity of the problem that can be estimated, and 10 by the type of customer, academic or commercial.[22] A number of these packages were also offered in "student" versions at substantially lower prices, segmenting the market even further. This evidence tends to confirm that in some cases, successful price discrimination is feasible and can be used to serve the academic market while covering some of the fixed costs via the commercial market.

Although price discrimination is widely used in these markets, it does have some drawbacks as a solution to the problem of software provision. The most important one is that features important to academics or even programs important to academics may fail to be provided or maintained in areas where there either a very small commercial market or no market, because their willingness to pay for them is much lower. Obviously this is not a consequence of price discrimination per se, but simply of low willingness to pay; the solution is not to eliminate price discrimination, but to recognize that PD production of some of these goods is inevitable. For example, a database of elementary particle data has been maintained by an international consortium of particle physicists for many years. Clearly such a database has little commercial market.

6 Conclusions

Among the activities that constitute academic research, the production of software and databases for research purposes is likely to be especially subject to underprovision and/or privatization. The reason is that like most research activities, the public goods nature of the output leads to free-riding, but that the usual norms and rewards of the "Republic of Science" are less available to their producer and maintainers, especially the latter. In this paper we presented a model that illustrates and formalizes these ideas and we used the model to show that the GPL can be a way to ensure provision of some of these goods, at least when the potential producers also want to consume them.

Although in this paper we have emphasized the beneficial role of the GPL as a coordination device for producing the public good, in these conclusions we also want to point out that the GPL is not a panacea that works in all situations, and one of those situations may indeed be the production of scientific software and databases. One reason is that in practice it is difficult to distinguish between the "upstream" activities, which, as we discussed, ought to be produced under public domain, and the "downstream" ones. As we noted in the paper, the latter may entail important complementary investments. Therefore, they could be more effectively conducted under private rules that enable the producers to raise the rents that are necessary to perform such investments. But the GPL "forces" the contributors to work under public domain rules. If one cannot properly distinguish between upstream and downstream activities, the downstream activities, with implied complementary investments, will also be subject to public domain rules. This makes it more difficult to raise the resources to make the investments, with implied lower quality of the product.

To return to the example of the introduction, the ***TeX*** User's Group reports the following on their website in answer to the FAQ "If ***TeX*** is so good, how come it's free?":

> It's free because Knuth chose to make it so. He is nevertheless apparently happy that others should earn money by selling TeX-based services and products. While several valuable TeX-related tools and packages are offered subject to restrictions imposed by the GNU General Public Licence ('Copyleft'), TeX itself is not subject to Copyleft. (http://www.tug.org)

Thus part of the reason for the spread of ***TeX*** and its use by a larger number of researchers than just those who are especially computer-oriented is the fact that the lead user chose not to use the GPL to enforce the public domain, enabling commercial suppliers of ***TeX*** to offer easy-to-use versions and customer support.

The so-called "lesser" GPL (LGPL) or other similar solutions can in part solve the problem. As discussed by Lerner and Tirole (2002), among others, the LGPL and analogous arrangements make the public domain requirement less stringent. They allow for the mixing of public and private codes or modules of the program. As a result, the outcome of the process is more likely to depend on the private incentives to make things private or public, and this might encourage the acquisition of rents in the downstream activities. But following the logic of our model, as we allow for some degree of privatization, the efficacy of the license as a coordination mechanism is likely to diminish. We defer to future research a more thorough assessment of this trade-off. Here, however, we want to note that when the importance of complementary investments is higher, one would expect LPGL to be socially more desirable. The benefits of having the downstream investments may offset the disadvantage of a reduced coordination in the production of the public good. By contrast, when such investments are less important, or the separation between upstream and downstream activities can be made more clearly (and hence one can focus the GPL only on the former), a full GPL system is likely to be socially better.

References

Cohen, W. M., R. Florida, and L. Randazzese, 2004, For knowledge and profit: university-industry research centres in the United States, Oxford University Press, Oxford, forthcoming.

Cohen, W. M., R. Florida, L. Randazzese, and J. Walsh, 1998, Industry and the academy: uneasy partners in the cause of technological advance, in: Noll, R. (Editor), The Future of the Research University, Brookings Institution Press, Washington, D. C.

Collins, S., and H. Wakoh, 1999, Universities and technology transfer in Japan: Recent reforms in historical perspective, University of Washington and Kanagawa Industrial Technology Research Institute, Japan.

Dalle, J.-M., 2003, Open source technology transfer, paper presented to the Third EPIP Conference, Maastricht, The Netherlands, November 22/23, 2003.

Dasgupta, P., and P. A. David, 1994, Toward a new economics of science, Research Policy 23, 487-521.

David, P. A., 2004, A tragedy of the public knowledge 'commons'? Global science, intellectual property and the digital technology boomerang, this issue.

David, P. A., 2002, The economic logic of open science and the balance between private property rights and the public domain in scientific data and information: a primer, National Research Council Symposium on The Role of the Public Domain in Scientific and Technical Data and Information (Washington, D.C.: National Academy of Sciences).

European Commission, 2002, Draft directive on the patentability of computer-implemented inventions (20 February), available at http://www.europa.eu.int/comm/internal_market/en/indprop/comp/index.htm

Foray, D., 2003, Innovation and knowledge openness: anatomy of the "private-collective" innovation model, presentation to the European Summer School in Industrial Dynamics, Cargese, Corsica

Gambardella, A., and B. H. Hall, 2003, On the stability of intellectual property regimes: IP protection versus open source/open science, Scuola Sant'anna Superiore Pisa and University of California at Berkeley:

Guena, A., and L. Nesta, 2004, University patenting and its effects on academic research: the emerging European evidence, this issue.

Hall, B. H., 2004, On copyright and patent protection for software and databases: a tale of two worlds, in O. Granstrand (Editor), Economics, Law, and Intellectual Property, (Boston/Dordrecht: Kluwer Academic Publishers), forthcoming.

Hall, B. H., A. N. Link, and J. T. Scott, 2001, Barriers inhibiting industry from partnering with universities, Journal of Technology Transfer 26, 87-98.

Harhoff, D., J. Henkel, and E. von Hippel, 2003, Profiting from voluntary information spillovers: How users benefit by freely revealing their innovations, Research Policy.

Hertzfeld, H. R., A. N. Link, and N. S. Vonortas, 2004, Intellectual property protection mechanisms and research partnerships, this issue.

Jensen, R., and Thursby, M., 1998, Proofs and prototypes for sale: The licensing of university inventions, American Economic Review, 91, 240-259.

Knuth, D. E., 1997, The Art of Computer Programming, Volumes I-III, Third Edition (Addison-Wesley, Reading, Massachusetts).

Kuan, J., 2002, Open source software as lead user's make or buy decision: a study of open and closed source quality, Stanford University.

Lerner, J. and J. Tirole, 2002, Some simple economics of open source, Journal of Industrial Economics L, 197-234.

Levinthal, D. and J. G. March, 1993, The myopia of learning, Strategic Management Journal 14, 95-112

Maurer, S. M., 2002, Promoting and disseminating knowledge: the public/private interface, Working Paper (September).

Nelson, R. R., 1959, The simple economics of basic scientific research, Journal of Political Economy 77, 297-306.

Nelson, R.. R., 2002, The market economy, and the republic of science (Columbia University, New York).

Olson, M., 1971, Logic of Collective Action: Public Goods and the Theory of Groups (Harvard University Press, Cambridge, Massachusetts).

Raymond, E. S., 1999, The Cathedral and the Bazaar (Sebastopol, California, O'Reilly).

Renfro, C. G., 2003a, A compendium of existing econometric software packages, Journal of Economics and Social Measurement 29, forthcoming.

Renfro, C. G., 2003b, Econometric software: the first fifty years in perspective, Journal of Economics and Social Measurement 29, 1-51.

TeX User's Group website, http://www.tug.org

Thursby, J., R. Jensen, and M.C. Thursby, 2001, Objectives, characteristics and outcomes of university licensing: A survey of major U.S. universities, Journal of Technology Transfer 26, 59-72.

Von Hippel, E, and G. von Krogh, 2003, Open source software and the private-collective innovation model: issues for organization science, Organization Science, forthcoming.

Von Hippel, E., 1988, The Sources of Innovation (Oxford: Oxford University Press).

Table 1: Econometric Software Packages

Type of seed funding	**Total number of products**	**Number commercialized**	**Average lag to commercialization**	**Average date of introduction**
research grants or own research	20	16	9.4	1979
quasi-governmental organization	5	4	16.4	1974
private (for profit)	5	5	0.8	1984
Total or average	**30**	**25**	**9.0**	**1979**

Table 2: Comparing Non-commercial and Commercial software

Features	Share of non-commercial	Share of commercial
Full windows, icons, menus interface (WIMP)	60%	60%
Interactive use possible	60%	68%
Macro files can be executed	40%	84%
Manipulate objects with icons/menus	60%	88%
Generate interactive commands with icons/menus	20%	60%

Table 3: Price Discrimination in Econometric Software

Price discriminate?	**No. of packages**
by size or complexity	3
academic/commercial	10
no discrimination	2
NA	2
sold as a service	8
free	5
Total	30

Figure 1: Equilibrium

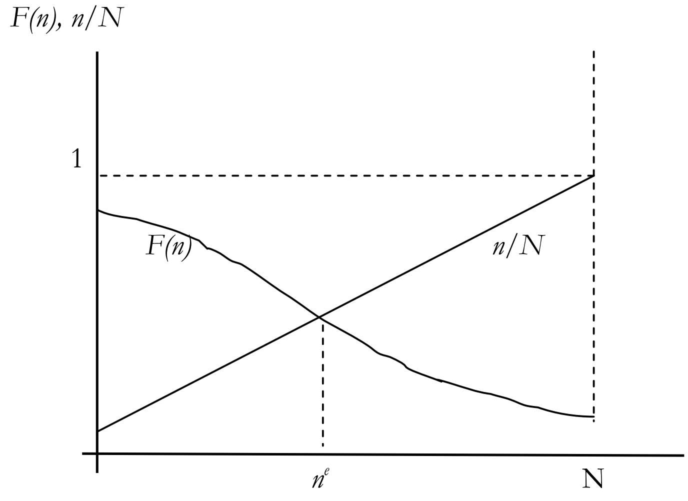
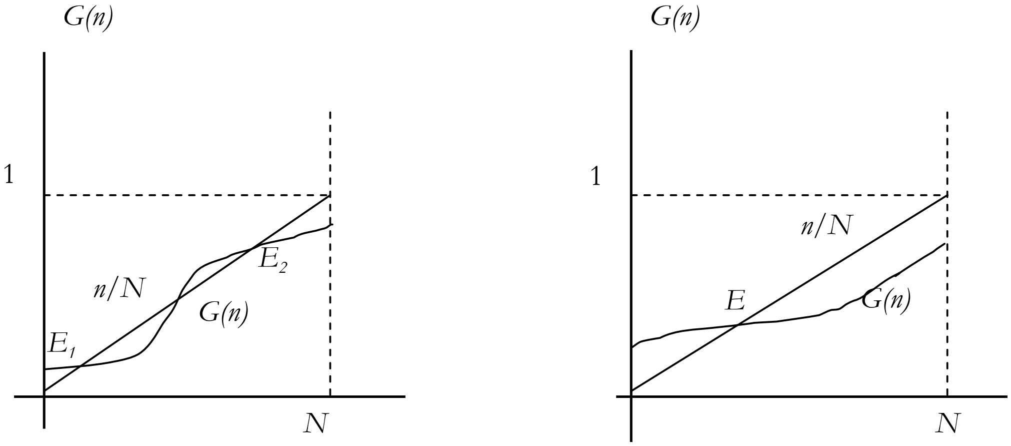


Figure 2: Equilibria



Two stable equilibria (E$_1$ and E$_2$)

One stable equilibrium (E)

[1] Conversations with Paul David on this topic have helped greatly in clarifying the issues and problem. Both authors acknowledge his contribution with gratitude; any remaining errors and inconsistencies are entirely our responsibility. We are also grateful to Jennifer Kuan for bringing some of the open source literature to our attention.

[2] See Hall (2004) for further discussion of problems on the boundary between public science and IP regimes.

[4] This brief history of ***TeX*** is drawn from the ***TeX*** User's Group website, http://www.tug.org. In giving a simplified overview, we have omitted the role played by useful programs based on ***TeX*** such as ***LaTeX***, etc. See the website for more information.

[5] WYSIWYG is a widely used acronym in computer programming design that stands for "What You See Is What You Get".

[6] For a more elaborate version of the model, see Gambardella and Hall (2004).

[7] See also Foray (2003) who has recently noted the inherent instability of systems of public domain/collective inventions when they are not based on stable property rights.

[9] An implicit assumption is that these ν researchers are those with π large enough so that each one will want to switch if all of them do. That is, the coordinator is indifferent as to the identities of the group of ν switchers.

[10] We show later that under the assumptions of our model, choice c) is not a real choice; if the project is launched under PD, it will have a GPL attached, provided adequate enforcement is available.

[11] This can thought of as the utility from the least rewarding project in the utility function (1). Thus, B is for example the minimum across all the projects i of the researcher of the maximum between $X_i(n_i)$ and $X_i(n_i) + \pi_i$, where n_i is the number of researchers working under PD in each field.

[12] When G increases faster with n, any increase in n from the intersection point implies that the share of researchers with $B \leq X(n)$ becomes higher than n/N, which induces more people to deviate. Also, many equilibrium configuration are possible, including $n=N$ and $n=0$. (In the latter case, e.g., $G(n)$ starts from zero and lies entirely below n/N; in the former, it lies entirely above n/N). Finally, we could have assumed that the researchers who do not work on the project benefit from the new project anyway. This implies that the condition to participate is $B \leq X(n) - \gamma X(n\text{-}1)$ where $0 \leq \gamma \leq 1$ is a parameter that measures

the impact of the new project when the individual does not participate. It turns out that the qualitative discussion of this section would not change, although of course the number of participants for any given project would be reduced.

[13] Note that $G(\cdot)$ defined earlier is then the marginal probability distribution of $\Gamma(\cdot)$.

[14] Clearly, correlation between B and π or the lack thereof can affect the participation in the project and not just the share of *PD vs. PR* participations.

[15] This argument should be familiar as it is the same as the argument used by some to justify Bayh-Dole and the granting of exclusive licenses for development by universities.

[16] The usual commercial web-based provision of data is based on a model where the user constructs queries to access individual items in the database, like looking up a single word in the dictionary. The pricing of such access reflects this design and is ill-suited (i.e., very costly) for researcher use in the case where research involves studying the overall structure of the data.

[17] This can be a real cost. The U.S. Patent Office, which provides a large patent database free to the public at large on its web server, has a notice prominently posted on the website saying that use of automated scripts to access large amounts of this data is prohibited and will be shut down, because of the negative impact this has on live individuals making queries.

[18] Another type of academic information product deserves mention here, academic journals. The private sector producers of these journals fact the same type of cost structure and have pursued a price discrimination strategy for many years, discriminating between library and personal use, and also among the income levels of the purchasers in some cases, where income level is proxied by country of origin.

[19] In principle, in the aftermath of the (1981) Diamond v. Diehr decision, patent protection might also be available for some features of econometric software. In this area, as in many other software areas, there is tremendous resistance to this idea on the part of existing players, perhaps because they are well aware of the nightmare that might ensue if patent offices were unacquainted with prior art in econometrics (as is no doubt currently the case).

[20] Unfortunately, it is not possible to identify precisely the nature of the seed money support for many of the packages from the histories supplied in Renfro (2003a), other than the simple fact that the development took place at a university.

[21] This quotation is from Hermann Bierens' website at http://econ.la.psu.edu/~hbierens/EASYREG.HTM

[22] The average ratio of commercial to academic price was 1.7. Assuming an iso-elastic demand curve with elasticity η and letting s=share of commercial (high demand) customers, one can perform some very rough computations using the relationship $\Delta Q/Q = -\eta \, \Delta P/P$ or $(1-s) = \eta \, 0.7$. If $\eta=1$, then the implied share of academic customers is 70 per cent. If the share of academic customers is only 30 per cent, then the implied demand elasticity is about 0.42.