Karlo Kauko

The Microeconomics of Innovation:

Oligopoly Theoretic Analyses with Applications to Banking and Patenting

Bank of Finland Studies
E:18 • 2000

The Microeconomics of Innovation:

Oligopoly Theoretic Analyses with Applications to Banking and Patenting

By
KARLO KAUKO
Lic Sc (Econ)

A doctoral dissertation to be presented, by permission of the Council for Academic Affairs of the Helsinki School of Economics and Business Administration, for public examination in the Nokia B200 auditorium, Runeberginkatu 14–16, Helsinki, at 12 noon on 12 May 2000.

Helsinki 2000

Karlo Kauko

The Microeconomics of Innovation:

Oligopoly Theoretic Analyses with Applications to Banking and Patenting

Bank of Finland Studies
E:18 • 2000

The views expressed in this study are those of the author and do not necessarily reflect the views of the Bank of Finland.

ISBN 951-686-651-4
ISSN 1238-1691
(print)

ISBN 951-686-652-2
ISSN 1456-5951
(online)

Gummerus Kirjapaino Oy
Jyväskylä 2000

(Published also as A-166, Helsinki School of Economics and Business Administration, ISBN 951-791-442-3, ISSN 1237-556X)

Abstract

The innovation activities of companies has long been a topic of interest in economics. Game theory models of oligopoly have since the start of the 1980s played a central role in the economics of innovation. In this study three game theory duopoly models are presented and each is used to analyse the firm's R&D activities.

The first model is used to examine the variables that affect the incentives of banks providing payment services to develop an interbank payment system. A customer of a large bank may be in an advantageous situation in that most of his payments will be effected in that bank's internal payment system, which is more reliable and otherwise superior to the interbank system. A key result derived from the model is that provision of payment services free of charge to customers often results in a distortion of banks' incentives to develop the system. A smaller bank will overinvest in the system in order to improve its relative competitive position. Because system improvement would only weaken the large bank's superior position, it will not have a strong incentive to improve the system. Since only one of the model's two banks is investing in the quality of the system, the investments will generally not be cost-effective. If fees are charged for payment services, the distortions in incentives are less serious, even though it is often the case that both banks overinvest in the sytem. When model results are compared to historical situations regarding payment systems, a number of consistencies are found.

The second model deals with the possibilities of a national government to influence domestic companies' investments in product development via patent laws that discriminate against foreign companies. If two countries have discriminatory patent laws in order to promote domestic companies' investments in product development, the results may well turn out to be offsetting. If just one of the two countries discriminates against foreign patent applicants, this may result in either more or less R&D effort by domestic companies, depending on the situation.

The third model is used to study patenting decisions by a company that has made an innovation. A company can monopolize its innovation by either patenting it or keeping it secret. Patenting is the only viable option if a competitor independently comes up with the same innovation. A patent application, by contrast, is a public document, the contents of which are useful to others who would like to develop substitute products. Patenting is thus not advantageous unless the competitor is likely to come up with the same innovation

independently. This means that a company will be the more inclined to patent an innovation, the more its rival invests in R&D. A risk-averse company is more inclined to patent than a risk-neutral one. This model is generally supported by empirical findings.

Keywords: innovation, oligopoly, banking, patenting

Tiivistelmä

Yritysten innovaatioita on jo pitkään käsitelty taloustieteissä. Peliteoreettiset oligopolimallit ovat 1980-luvun alusta lähtien olleet keskeinen osa innovaatioiden taloustiedettä. Tässä tutkimuksessa esitetään kolme peliteoreettista duopolimallia, joista kukin analysoi yhtä yritysten tutkimus- ja kehitystoimintaan liittyvää kysymystä.

Ensimmäisessä mallissa analysoidaan tekijöitä, jotka vaikuttavat maksuliikennepalveluja tarjoavien pankkien kannustimiin kehittää pankkien välistä maksujärjestelmää. Suuremman pankin asiakkuus on usein houkutteleva vaihtoehto, sillä suuren pankin asiakkaat käyttävät etupäässä pankin sisäistä maksujärjestelmää, joka on parempi ja luotettavampi kuin pankkien välinen maksujärjestelmä. Yksi mallin keskeisistä johtopäätöksistä on, että jos maksupalvelut ovat asiakkaille ilmaisia, pankkien kannustimet kehittää järjestelmää ovat usein vinoutuneet. Pieni pankki yli-investoi järjestelmään parantaakseen suhteellista kilpailuasemaansa. Järjestelmän parantaminen heikentäisi suuren pankin suhteellista paremmuutta, joten suuri pankki on haluton kehittämään järjestelmää. Koska vain toinen pankeista investoi, tehdyt investoinnit parantavat järjestelmän laatua kustannustehottomasti. Jos maksupalvelut eivät ole asiakkaalle ilmaisia, vinoumat pankkien kannustimissa ovat vähäisemmät, joskin huomattavan usein niin pienet kuin suuretkin pankit jossain määrin yli-investoivat järjestelmään. Mallin tulemia verrataan Suomen maksujärjestelmien historiaan, ja tiettyjä yhtäläisyyksiä on havaittavissa.

Toinen malli käsittelee kansallisen hallituksen mahdollisuuksia vaikuttaa ulkomaalaisia syrjivän patenttilainsäädännön avulla kotimaisten yritysten tuotekehitysinvestointeihin. Jos kaksi valtiota soveltaa syrjiviä säädöksiä edistääkseen kotimaisen yrityksen tuotekehitysinvestointeja, säädösten vaikutukset helposti kumoavat toisensa. Jos vain toinen valtioista syrjii ulkomaisia patentinhakijoita, kotimaisten yritysten tutkimus- ja kehitystoiminnan määrä tapauksen mukaan joko kasvaa tai vähenee syrjinnän antaman suojan vuoksi.

Kolmas malli käsittelee innovaation tehneen yrityksen patentointipäätöstä. Yritys voi monopolisoida keksintönsä joko patentoimalla sen tai pitämällä sen salaisena. Patentointi on ainoa tapa sellaisessa tapauksessa, jossa kilpailija tekee itsenäisesti saman keksinnön. Toisaalta patenttihakemus on julkinen asiakirja, jonka sisältämät tiedot auttavat jäljittelijöitä kehittämään korvaavia tuotteita. Patentointi ei siis kannata, ellei kilpailija melko todennäköisesti tee itsenäisesti samaa keksintöä. Yritys on siis sitä halukkaampi patentoimaan keksintönsä, mitä enemmän sen kilpailija sijoittaa tutkimus- ja kehitystoimintaan.

Riskiaversiivinen yritys on halukkaampi patentoiman kuin riskineutraali. Malli saa tukea empiirisistä havainnoista.

Asiasanat: innovaatiot, oligopolit, pankit, patentointi

Foreword

My greatest indebtedness is to Professor Pekka Ilmakunnas. I have greatly benefited from his many insightful comments and general encouragement at various stages of the project.

Moreover, I am grateful to my official examiners, Professor Johan Willner and Dr Heli Koski. Their constructive criticism during the final stages of the project helped me to finalize the thesis.

This study took much longer than what was originally intended, and I am fully aware that I am unable to mention all the other persons who have contributed with their comments at the various stages. But I must at least mention the enormous help that I received in the form of comments and suggestions from Dr Juha Tarkka, Dr. Jouko Vilmunen, Professor Mihkel Tombak, Professor Rune Stenbacka, Professor Oz Shy, Mrs Vappu Ikonen, Mr Jukka Vesala and Professor Trond Olsen. I should also mention that the continuous encouragement that I received from my superiors, Dr Heikki Koskenkylä and Dr Markku Malkamäki of the Bank of Finland's Financial Markets Department, was essential to my perseverance.

Financial support provided in the early stages of the project by the Yrjö Jahnsson Foundation, Nordisk Forskerutdanningsakademi, Eevi and Eemil Tanninen Foundation and the Helsinki School of Economics and Business Administration Support Foundation is gratefully acknowledged. I should like to thank the research department of the Bank of Finland, The Helsinki School of Economics and the Norwegian Research Centre in Organization and Management for providing me with the necessary physical facilities.

My thanks also go to Mr Glenn Harma who skilfully revised the language of this study, and to Mrs Päivi Lindqvist who helped to finalize the layout for the publication.

Last but not least, I would like to thank my wife Riitta Alamiekkaoja for her encouragement.

This study is dedicated to my children – Otso, Ilari and Ursula.

Helsinki, March 2000
Karlo Kauko

Contents

1 Introduction

1.1 Endogenous innovation – some general observations

1.1.1 Research and development – a key factor in economic development

At least in advanced countries, labour productivity has since the early days of industrial revolution had a particular property not shared by many other economic variables; it has a persistent, unidirectional trend. The capital to output ratio, the functional distribution of income, the strength of business cycles, and the real rate of interest may differ somewhat from their levels of, say, a hundred years ago, but one is inclined to say that the changes in these variables have not been dramatic.

Viewed over the long term, the increase in productivity has been spectacular. In Finland, the volume of output in manufacturing industries was 37 times higher in 1997 than in 1860, whereas the number of persons employed in manufacturing industries was only 6 times higher. Because of shorter workweeks, the number of hours worked must have risen by even less than that.[1]

Apparently there must have been a spectacular increase in total factor productivity. It would be completely unrealistic to argue that growth in the stock of capital goods alone could explain this increase in output.

The importance of technological factors as a source of this productivity growth has been one of the main subjects of the "residual debate". In the literature of this field, economic growth is explained with as many other factors as possible, and the residual is assumed to reflect technological progress. Even in the most detailed analyses, where as much as possible is taken into account, for instance in Jorgenson et al. (1987) and Denison (1985), a significant portion of output growth appears to be explained by technological progress, because no other explanation is found.

Not only is the quantity of output strongly affected by technological progress. In addition, the variety of final consumer

[1] Hjerppe 1985 and own calculations based on data collected by Statistics Finland.

goods is completely different from what it was three or four generations ago. One hundred years ago, not even the wealthiest consumers were able to buy pocket calculators. Nowadays, slide rules are sold only in antique shops. Very few manufactured consumer goods have not undergone any changes since, say, the end of the 19th century. Therefore, empirical studies that ignore the composition of output underestimate, rather than overestimate, the impact of technological change on the economy.

In addition, technological innovations can be essential background factors in the appearance and disappearance of economic institutions. According to Jutikkala (1953, p. 395), the invention of the railway was a key factor behind the (re) introduction of the concept of the limited liability company. Or, to take a modern example, some observers expect that electronic commerce over the Internet will have a profound and widespread impact on the economy, and it might even eliminate most traditional retail outlets.

Because economic growth is to a large extent determined by technological progress, it is not surprising that numerous empirical studies demonstrate that R&D efforts have an observable impact on productivity and output growth. This effect has been observed both at the industry level (Mansfield, 1988; Badulescu, 1988, p. 31; Gowdy, 1993; Perelman 1995) and at the macroeconomic level (Artus & Kaabi, 1993; Guellec & Ralle, 1993; Evenson et al., 1988, p. 27; Gittleman & Wolff, 1995)

Research and development as an economic factor becomes even more important if it is measured by the amount of resources devoted to R&D. The amount of resources allocated to R&D grows decade after decade at a far higher rate than the economy in general (Mansfield, 1969, p. 4–9; Grossman & Helpmamn, 1991b, p. 10 and Ferrantino, 1992, p. 689). In Finland, there are already several manufacturing industries that invest more in research and development than in tangible capital goods, and if the trend continues, the number of such industries will increase in the future.

1.1.2 The economics of endogenous innovation – a short history

Pre-Schumpeterian contributions

The concept of innovation as an endogenous economic factor appears in embryonic form in the writings of at least two 19th century economists. The Scotsman John Rae wrote as early as 1834 on the

psychology of the inventor and on the social benefits of promoting innovative activities (Heertje 1977, p. 81–82). In 1855, the German economist von Mangolt wrote about entrepreneurial profits as a reward to innovative efforts (Ekelund and Hébert, 1983, p. 281).

It is also possible to find the concept of endogenous innovation in the context of the German historical school. In Schmoller's book "Grundriß der Allgemeinen Volkswirtschaftslehre" (1904), technological development is regarded as the key determinant of economic development. Schmoller argued that technological progress is strongly affected by the economy (Rosegger, 1988).

Schumpeter

Endogenous innovation in the economics literature is often associated with Joseph Alois Schumpeter (1883–1950). According to Schumpeter's book "Theorie der wirtschaftlichen Entwicklung" (1912), the Walrasian approach is incapable of explaining economic development, except by ascribing it to exogenous changes in environmental factors (p. 108–118). The main thesis of the book, presented in the second chapter (p. 103–198), is that from the point of view of economic dynamics, price competition matters because entrepreneurs have a strong incentive to avoid it. The entrepreneur must introduce something new and unique in order to gain a temporary monopoly position. Economic development is the consequence of such new business ideas and technological innovations.

However, it took decades before theoretical studies on R&D became an established field of literature among other economists. "Theorie der wirtschaftlichen Entwicklung" was not translated into any language before the 1930s. By then, mainstream economics had evolved so that Schumpeter's methodology was no longer updated. For instance, it was no longer commonplace to combine sociology and economics. Instead, the use of mathematical models (which Schumpeter had not used) had become much more popular. This disparity between Schumpeter's book and mainstream economics may have contributed to a very curious situation. Even though Schumpeter is among the best known economists of the 20th century, not even Schumpeter's students – including Gottfried Haberler, Paul Samuelson and Paul Sweezy – are known to have tried to continue his work (Madarász, 1991, p. 219).

Schumpeter's own contributions "Theory of Business Cycles" (1939) and "Capitalism, Socialism and Democracy" (1942) were

strongly influenced by his own earlier ideas. "Capitalism, Socialism and Democracy" discusses how capitalism might lose its dynamic nature because innovation can be done mainly in large corporations in concentrated industries, which diminishes the possibilities of individual innovator-entrepreneurs. Although one of the book's main conclusions seems to have been wrong, the book is often considered to be Schumpeter's most important work.

Endogenous innovation in macroeconomics

In the economics of distribution, the concept of endogenous innovation appeared almost 70 years ago: Hicks (1963, p. 121–127, originally published in 1932) discussed the likely impact of factor prices on what he called "induced inventions", especially the impact of high wages on the emergence of labour saving production technologies. Later, especially in the 1960s, these issues aroused the interest of several economists who were studying income distribution (Heertje, 1977, p. 179–181).

In the late 1980s and early 1990s, endogenous innovation has been introduced in macroeconomics on a much larger scale than before. In the mid-1980s, Kirzner (1985) concluded that the lack of inventiveness was one of the main problems of growth theories. This critique is no longer justified. The new growth theory is to a large extent based on the concept of endogenous innovations. (See Romer, 1989; Grossman and Helpman, 1990; Young, 1991 and Aghion and Howitt, 1992.)

Endogenous innovation has been incorporated in trade theories as well. (Rivera-Batiz and Romer, 1991a, 1991b; Grossman and Helpman, 1991a, 1991b; Blackburn and Hung, 1993). Some preliminary attempts to have also been made endogenize the flow of innovations in real business cycle models (Stadler, 1990).

R&D in industrial organization literature

It would not be justified to say that endogenized technological change had become a central topic in any field of economics before the 1960s. In the industrial organization literature of the period, there were a lot of exploratory statistical analyses, often inspired by theoretical ideas that were not in line with strict orthodoxy. During the era of the "structure – conduct – performance" paradigm, there was no reason

not to include some R&D indicators as additional variables in the analysis whenever statistics on innovative efforts were available.

One of the most popular hypotheses to be tested was taken from Schumpeter's "Capitalism, Socialism and Democracy" (1942): market concentration and large firm size should encourage innovative efforts. This literature was reviewed by Kamien and Schwartz (1975). Since then, this topic seems to have become relatively less popular, and it has been questioned whether the whole approach has been meaningful (Culbertson, 1985 and Cohen and Klepper, 1992). Nevertheless, the topic has not been abandoned.

Testing the "Schumpeterian hypothesis" has not been the only topic that has inspired empirical work. Since the 1960s, econometricians have measured the statistical covariation between various indicators of corporate R&D efforts and numerous firm-specific factors. Studies on the covariation between R&D and export orientation may be some of the best examples of a research tradition that has provided a rather robust empirical fact: at the micro level, export orientation and innovations are positively correlated under a wide variety of circumstances (Zimmermann, 1987; Entorf, 1988; Schlegemilch, 1988; Zimmermann and Schwalbach, 1991; Zif et al., 1990, Glick, 1982; Ito and Puick, 1993; Hirsch and Bijaoui, 1985; Bregman et al., 1991; Leppälahti and Åkerblom, 1988, p. 28).

Possibly as a consequence of both the existence of these econometric papers and the breakthrough of game theory in the industrial organization literature, game theoretic analyses of R&D and endogenous innovation in oligopolistic industries became popular in this literature in the 1980s.

Fisher (1989) discusses the tendency of modern industrial organization literature to provide inconsistent results. Oligopoly theoretic studies on endogenous innovation are excellent examples of this tendency. Beath et al. (1989a) concluded that game theoretic models of patent races provide very few generally valid results. It is possible to find other examples of questions analysed with game theory models that provide completely different results. According to the model of Brander and Spencer (1983), R&D subsidies in an international duopoly would cause a transfer of wealth from the foreign rival to the subsidized domestic firm, thus increasing national welfare. However, Beath et al. (1989b) demonstrated that, with slightly different assumptions, it would be a better national strategy to tax R&D instead of subsidizing it. Firms are not willing to carry out as risky R&D projects as a planner would do (Dasgupta and Stiglitz, 1980), or they may have an incentive to favour projects that are too risky to maximize social welfare (Klette and de Meza, 1986).

Knowledge spillovers between firms may induce firms to limit their R&D efforts (Spence, 1984), or such spillovers may encourage them to increase their R&D efforts (Bondt et al. 1992 and Cohen and Levinthal, 1989).

This variety of opposing results does not necessarily imply that at least some of the models are "wrong", and should be abandoned. One model might be a good description of the situation in one industry, and another might suit a different industry. Therefore, the inconsistency of results probably implies that if one wants to analyse a given real life situation, say, for policy making purposes, he should not take just any model in the field. Careful attention should be paid to selecting the most suitable model.

It is often nearly impossible to determine the relevance of different models on purely theoretical grounds. Therefore, the predictive power of competing theories should be a very central criterion in assessing the applicability of different models in a given situation. Unfortunately, recent theories have not been a major source of inspiration for econometricians who have studied the determinants of technology efforts and/or innovative output. If one wants to evaluate the relevance of different theories about eg patent races, the existing empirical literature has little to offer. At least a partial explanation for this unfortunate situation may be that the results of theoretical research are not always testable. It would be an important challenge for theoreticians to try to construct models that would be testable, preferably with data that are readily available.

1.2 The role of government in technological development

Several arguments have been presented in favour of the assertion that different market failures are likely to arise when firms decide on their R&D budgets. In many cases, it has been concluded that firms typically under-invest in new technologies. These arguments are often related to the public good nature of knowledge, improved consumer welfare or spillover effects. Such arguments were presented already in the late 1950s and early 1960s (see Nelson, 1959; Arrow, 1963; Usher, 1964).

Fölster (1991) in his pragmatic study on Swedish policies lists potential reasons why a government could find it reasonable to try to encourage R&D in the private sector. Firms, especially minor ones, may be unwilling to engage in uncertain R&D projects that are large

in relation to the size of the company. From the point of view of the whole society, the uncertainty related to hundreds of independent projects is not a serious problem, because the risk is diversified away. In addition, there may be scale economies in R&D at the industry level: knowledge spillovers help rivals to utilize their R&D resources more efficiently. A firm would not normally invest anything in order to help its competitors.

In the 1980s and 1990s, many writers used game theoretic approaches that differed substantially from the classical analyses, focusing on rather simple positive externalities. These approaches are often related to imperfect competition and are based on game theory. A seminal contribution in this field is Dasgupta and Stiglitz (1980). Strictly speaking, Dasgupta and Stiglitz did not write about technology policies, although they compared competitive market outcomes and monopoly practices with the social planner's optimal choices. Their contribution presents a lot of potential market failures that can emerge in the allocation of resources to R&D. It is difficult to briefly summarize the findings of this paper because the authors reach several rather different conclusions. They analyse such issues as the number of firms or laboratories engaged in the activity, the amount of resources devoted to R&D and the speed of research. One of the key assumptions in their paper is that the first to innovate can patent the invention, which makes the speed of R&D in competitive situations higher than what would be socially optimal. In addition, the number of firms engaged in R&D can be larger than what would be desirable, with each of the companies may doing too little research. The paper of Dasgupta and Stiglitz has already become a classic. Many authors have referred to it, though it might not be entirely justified to argue that the authors have established a school of theoretical research.

Patent race models have become a classical sub-field of the game theoretic literature on R&D. In many cases, patent races lead to levels of R&D investment that exceed the social optimum. This research tradition will be discussed in section 1.3.

Moreover, it is possible to find contributions that are not pure patent race models. For instance, Romano (1989) analyses optimal innovation policies in two different situations. A firm that is simultaneously both the single potential R&D performer and one of the actual sellers can invest in R&D, or alternatively a great number of firms can. If successful, R&D lowers production costs. If patent life is finite, it is always justifiable to subsidize the research monopoly. After the patent has expired, perfect competition prevails and the social benefits of the invention are still large, but the inventor gets nothing. In the case of numerous firms competing for the patent, there may

instead be excessive research already in the market outcome, and an R&D tax would be more appropriate than subsidies.

Technology investments in mixed oligopolies have also been studied. Delbono & Denicolò (1993) analyse the situation for a mixed oligopoly, where a public firm could be an efficient policy tool to limit excessive R&D in a technology race where the winner-takes-all principle obtains. A public firm would invest less in R&D than a private firm and would thus discourage the private firm from investing excessively. This would increase social welfare.

Stoneman & Diedersen (1994) argue that there are probably rather serious market failures related to the diffusion of new technologies, and they emphasize the potential benefits of policies aimed at intensifying the diffusion process. Metcalfe (1994) discusses why and how technology policies should also cover the field of technology selection, ie the choice between known technologies.

One can easily make an interesting observation concerning real life technology policies and these sophisticated models. Many contributions imply that in non-regulated oligopolistic industries, excessive investments in R&D are likely to take place. However, it seems that in almost all cases, policy-makers try to promote innovative activities. The possibility that the government should, say, impose some extra taxes on research laboratories in order to deter excessive investment in R&D is normally not even discussed among politicians and civil servants. For instance, in OECD (1998) there is a detailed discussion on the government's contribution to technological progress, but the possibility that a national government would try to discourage the private sector from spending too much on R&D is not even mentioned.

In addition, there are several contributions that discuss the choice of optimal technology policies. Stoneman (1987, p. 4) defines the concept "technology policy" as a set of policies with the intent of affecting the process of technological innovation. Enforcing patent rights is a typical technology policy measure. It would also be completely reasonable to include such factors as the educational system, basic academic research carried out in universities and the behaviour of the government as a purchaser of technologically advanced goods.

Rothwell and Zegveld (1981) attempted to present a typology for direct and indirect policy tools that a government can use to affect the flow of successfully commercialized innovations. They make a distinction between factors affecting the supply of new inventions (technology push policies) and the factors affecting the demand for new products and processes (technology pull policies). The

government could try to boost the supply of inventions by improving both the availability of R&D inputs, such as qualified manpower, as well as firms' access to funding. As to the demand for innovations, the government could purchase domestically produced technology-intensive items or sponsor the adoption of new products and processes.

Technology policies used by different governments have often been reviewed and evaluated; for instance, the Japanese system has been analysed in Freeman (1987) and Fransman (1995), and the Swedish system by Fölster (1991). Oberender and Fricke (1993) deals with the EU system. Berg et al. (1996) discusses the prospects for future Estonian technology policy. According to Ergas (1987), some large OECD countries (US, France, UK) have subsidized, above all, large high-tech projects, whereas some other countries have focused on the diffusion of existing innovations and small firms (West Germany, Sweden, Netherlands).

1.3 The patent system in the previous literature

Patent protection has probably been used for a much longer time than any other innovation policy tool (See eg David and Olsen, 1992).

The strengths and weaknesses of the patent system have often been compared to other possible technology policy tools. Wright (1983) compares, in an imperfect information context, patents and a prize that could be shared between successful innovators. Patents can lead to resource-wasting technology races. On the other hand, imperfect information may make them a better alternative; an insufficiently informed government cannot set an optimal prize, but well-informed firms are aware of the costs and benefits. Romano (1991) analyses the problems of both patents and subsidies; patents lead to monopoly pricing. Subsidies have to be financed with taxation, and taxes may lead to resource misallocation. Whether patents lead to a better allocation of resources than public funding or vice versa depends on the case. In many cases, the best possibility seems to be a combination of these policies.

Many studies have been done one the impact of the patent system as such on firms' incentives to carry out R&D. The worthiness of the outcome has often been analysed in these studies. A central finding of the "classical" patent race literature is that competing firms end up in a prisoner's dilemma situation. Excessive investment in R&D takes place when firms try to win the race. By increasing its own R&D

effort, a firm reduces others' possibilities of winning the race, which is a negative externality. On the other hand, due to time preference, consumers and would-be licensees are better off if firms speed up the process. These effects were discussed already by Loury (1979), Dasgupta and Stiglitz (1980) and Reingaum (1982). These early models were based on a "memoryless" process, where the possibilities of a firm to invent depend solely on its current R&D expenditure. Past efforts are irrelevant. Subsequently, some models have been based on the idea that the possibilities to succeed depend on accumulated experience. Fudenberg et al. (1983) and Harris and Vickers (1985) demonstrated that in such races the weaker firm might voluntarily give up.

Beath et al. (1989a) concluded that these game theoretic models of patent races provide very few generally valid results. This conclusion still seems to hold.

It is also possible to find completely different theoretical contributions on the welfare effects of the patent system. For instance, detailed studies have been done on how patent legislation could affect technological progress. Scotchmer (1991) discusses the optimal degree of patent protection in different cases. If inventions are based on previous inventions, excessive patent protection may deter technological progress by impeding others from further developing patented technologies. How broad patent protection should be may depend on whether collusive licensing and R&D joint ventures are allowed.

Theoretical work on the patent system has also been combined with international economics. Aoki and Prusa (1993) introduced an entirely new topic by analysing how a national patent authority could promote domestic R&D by discriminating against foreign innovators who compete with domestic firms. Adams (1998) extended this analysis and concluded that discrimination against foreign imitators could encourage domestic R&D in the case of infant industries whereas, in mature industries, it probably has the reverse effect. Marjit and Beladi (1998) analysed whether it is reasonable for a government of a developing country to introduce product patents if the foreign patent holder might deter local production of cheaper variants that would be affordable to consumers of the low-income country.

Most of the existing literature is based on the idea that the patent system can be socially desirable only because it gives firms incentives to carry out R&D. However, patents may be beneficial to social welfare even if it is assumed that inventions are exogenous. In the presence of learning-by-doing effects and related spillovers, finite life patents might be socially desirable because they can give the patentee

an incentive to exploit the invention at a socially desirable rate (David and Olsen, 1992).

In most of these contributions, it is assumed that it is always essential to patent the invention, because it is the only available way to monopolize the results. Thus the relevance of this literature is largely dependent on whether other means, such as secrecy, would be more efficient.

In the empirical literature, the impact of the patent system on innovative activities is a surprisingly seldom studied topic. This is to some extent understandable, because it is quite difficult to imagine how the issue could be studied with econometric analyses based on observed firm behaviour. Almost all firms have had the possibility to apply for patents, and therefore there is no control group of firms not having the patenting option. Nevertheless, in the light of the existing interview studies, the impact seems to be relatively weak (Mansfield, 1986; Levin et al., 1987; Harabi, 1992 – reviewed in Franke, 1993).

The efficiency of patent protection from the point of view of a patentee has been analysed. Lanjouw (1998) presented some estimations of the value of patent protection in West Germany over the period 1953–1988. Patentees have to pay renewal fees and legal expenses in order to keep their patents valid and to deter imitations. In the light of the willingness to pay these fees and costs, the patent system has generated in Germany an aggregate value worth about 10 % of R&D costs.

It is possible to find a few game theoretic contributions that pay attention to the "to patent" dilemma. Unfortunately these contributions do not contain any empirical sections. Moreover, there are several empirical papers where the covariation between patents and R&D effort is measured. These contributions will be described in chapter 4.

1.4 Technology and banking in the previous literature

Even though banks were among the first institutions to acquire computers in the 1950s, banking has not traditionally been classified as a technology-intensive industry. Alhonsuo and Tarkka (1989) estimated that productivity growth in Finnish banking was stagnant for lengthy periods of time. Since the 1960s, the development of both total factor productivity and labour productivity had been much weaker in banking than in other service industries and manufacturing. In the late 1970s, productivity took off, and in the 1980s productivity

grew faster in banking than in manufacturing. Interestingly, the productivity take-off in banking in the early 1980s took place simultaneously with the launching of ATMs and debit cards. Thus, it might be realistic to argue that this productivity take-off was at least partly caused by technological change, even though deregulation has probably also contributed to increasing productivity.

Banking technology consists, to a large extent, of information technology, including software and hardware. Very few banks have developed any hardware equipment themselves. Instead, many of them have spent large sums in both developing own software applications and in adapting and installing externally purchased hardware. Both of these roads to new technologies are dependent on the resources the bank can allocate to information technologies. Frei et al. (1997) observed that size of information technology staff has a strong and statistically significant impact on output as measured as the sum of deposits and loans.

Technological progress probably reduces average costs in banking, even though the effect is surprisingly weak. For instance, Karafolas and Mantakas (1996) failed to find any evidence to support the hypothesis that Greek banks had become more cost efficient during the period 1979–1989, even though a number of technological innovations had been introduced. There may be at least two reasons for this.

1) To a large extent, innovations are related to the basic infrastructure. When ATMs were introduced, the resulting new distribution chain did not replace the branch network. Instead, it became an entirely new network with its own fixed costs.
2) Innovations may affect customers' behaviour in such a way that banks' costs are affected. For instance, large ATM networks may encourage customers to make more, but smaller on average, cash withdrawals. The average cost of a transaction would then diminish, but the number of transactions would increase and total costs to the bank could increase. This hypothesis is supported by US (Berger, 1985) and Spanish (Maudos, 1995) data.

Production functions in banking have been the subject of several empirical contributions, and these studies already form an established research tradition. Measuring output in banking is not as straightforward as in manufacturing. There are two different ways to operationalize this concept: the intermediation approach and the production approach. In the intermediation approach, a combination of different balance sheet items is used as a proxy for output, whereas

with the production approach, output is measured by number of accounts or other indicator of activity (Humphrey, 1985).

Another major problem with this literature is due to the fact that many (most) banks produce several different services simultaneously. A traditional way to handle this problem is to use the so-called translog production function, where the logarithmic total cost of a bank is explained with logarithms of different outputs and input prices. However, the applicability of this model is highly questionable in cases where certain banks do not produce all services (Noulas et al., 1993).

Eventual economies of scale have been a central topic in this production function literature. It is difficult to conclude what these studies have revealed, because the results are, to a large extent, inconsistent. It has often been concluded that there would be clear economies of scale among small and medium-sized banks but not among large ones. This conclusion is corroborated at least by Pulley and Braunstein (1992) and Esho and Sharpe (1995, p. 1151). Nevertheless, it is possible to find contrary evidence. Among European (Altunbas and Molyneux, 1996) and Japanese (McKillop et al., 1996) banks, average costs are a declining function of bank size, even among very large banks.

There is some evidence concerning the impact of technological progress on scale economies, but the results are mixed. Among Spanish savings banks technological progress has improved efficiency least among the smallest banks. Scale economies in the maintenance of an ATM network are strengthened via technological progress, whereas scale economies in lending did not undergo major changes (Maudos, 1995). Among small Bavarian local banks, by contrast, technological progress improved cost efficiency above all among very small institutions (Lang and Welzel, 1996). In the 1980s, US banks adapted new technologies characterized by economies of scale that were weaker than before (Beard et al., 1997).

In addition to the impact of technological progress on economies of scale, the propensity of banks to adopt new technologies has been analysed empirically. A traditional hypothesis in the economics of technology has been the so-called “Schumpeterian hypothesis”, according to which firms in concentrated industries are interested in innovating, because there are few competitors who might imitate. This hypothesis is consistent with the observation reported by Hannan and McDowell (1987) that banks in concentrated local markets adopted ATMs earlier than banks in highly competitive markets. Moreover, US banks with previous heavy investments in technology and intense

inter-organizational relationships were among the first to adopt video banking (Pennings and Harianto, 1992).

To sum up, there are many empirical contributions concerning technologies in banking. The two main topics have been production functions and the incentives of banks to acquire new technologies.

Because banking has not always been classified as a technology-intensive industry, it is not surprising that few models concerning endogenous technologies in banking were presented before the 1980s. In recent years, some theoretical contributions that try to focus on the specifics of technology competition in banking have been presented. The main subjects to be analysed have been spatial differentiation and remote banking. As with many other theoretical papers concerning endogenous innovations, these contributions are often based on game theoretic analyses.

Bouckaert & Degryse (1995) have presented reasons why banks might be reluctant to adopt new forms of customer service, such as phone banking. If interest rates offered to the public are strategic complements, banks might not have adequate incentives to introduce new forms of customer service. New customer service channels would reduce customers' transaction costs and thus weaken the possibilities to use market power and thereby intensify interest rate competition. Strategically-behaving banks would try to avoid this.

Degryse (1996) presented a game theoretic duopoly model concerning remote banking in a spatially differentiated market. Banks may offer remote banking services, if they prefer to do so. If they do, the importance of spatial differentiation diminishes. The incentives of banks to offer such services depend strongly on customers' preferences. If customers are not particularly willing to use remote banking services, banks have little incentive to invest in the service. The investment would above all intensify price competition, but would not give the investor much advantage in the struggle for market share. If at least some customers always prefer remote services, banks have an incentive to invest heavily in developing the quality of this service.

Vesala (1998) presents a model of banking competition, in which diffusion of electronic banking and strengthening of nonbank competition for savings are studied as factors that diminish the benefits of branch and ATM networks. Remote banking intensifies price competition and reduces difference in loan and deposit rates across banks. Moreover, it reduces the optimal numbers of branches and ATMs. Competition increases permanently unless banks are able to redifferentiate from rivals through novel innovation that compensates for the reduced value of network differentiation.

Capacity collusion is shown to reduce the sizes of branch and ATM networks as well as banks' mark-ups in loan and deposit rates. ATM compatibility reduces the total number of machines and under certain conditions raises deposit rates.

A central limitation of these theoretical analyses is their strong emphasis on connections between customer and bank, such as distribution networks and remote banking. Indeed, these models provide us with few if any insights concerning banks' internal processes, let alone processes between different banks. In real life, many banking technology inventions are related, above all, to payment services rather than to deposit collection or lending.

1.5 Outline and purpose of the study

This work contains three rather independent models, one of them focusing on banking and the other two on patenting.

In previous microeconomic literature concerning banking, there are several models that analyse lending and borrowing under asymmetric information. This topic is certainly relevant, but banks engage in other activities as well. One of the key functions of a modern banking system is the payment system. As we shall see in the section 2.6, economic history has demonstrated that banks' ability to provide their customers with payment services can have a substantial impact on their market shares. Nevertheless, it is difficult to find any theoretical models that help to analyse the specifics of payment services as a product of the banking industry. Chapter 2 hopefully contributes to our knowledge about this little studied topic.

In the model, two banks compete for customers in the market. The focus is on the quality of the interbank payment system and its impact on customers' choices in a Hotelling duopoly. If the interbank payment system functions poorly, sending/receiving interbank payments is slow and transactions can fail, implying that customers would have good reason to avoid such transactions. Customers would tend to prefer to use the same bank as the majority of other customers because, by definition, most potential counterparties in payment transactions use the bigger bank. Thus, the problems of interbank payments could be largely avoided by using the same bank as most payment transaction counterparties. Hence, from the point of view of the small bank, a poor interbank payment system is an obvious threat to its market share. By investing in the system, the small bank could reduce this competitive disadvantage. The large bank has incentives to

invest in the system if and only if the investment helps it to collect more payment service fees.

Central banks in different parts of the world have been more or less involved in designing, developing and administering interbank payment systems. There are several theoretical contributions that analyse the role of the central bank as a key institution of the payment system, but the focus of these has normally been on the design of an optimal clearing system. However, it is difficult to find any studies that present theoretical insights related to the optimal degree of central bank involvement in developing the interbank payment system. Should the banking industry be left alone to develop the kind of system it prefers, or should the central bank take the leading role in developing the system? What kinds of factors affect the optimal central bank policy in this area? Chapter 2 attempts to shed light on these policy questions.

As a rule, central bank intervention seems to be particularly important if banks either cannot price their payment services or if they voluntarily decide not to charge fees. In these cases, the non-existent fee revenue from payment services obviously cannot be the incentive for the private sector to invest. Instead, payment services are a tool in the struggle for market share, and it turns out that optimal private investment in the system is at its maximum when the societal need for the system is most limited, and *vice versa*. Thus, active central bank involvement is essential. If banks do charge fees for making interbank payments, their incentives are less distorted, and central bank involvement is not as essential.

Chapter 3 analyses another topic that seems to have been analysed by relatively few authors. In its analysis of the impact of protectionist policies on welfare, the existing theoretical literature focuses excessively on tariff protection. Tariff protection is certainly an important topic, but there are now other topics worthy of attention. Moreover, tariff protection is no longer as commonplace as it used to be. Knowledge and information are now becoming more and more important for companies, both as factors of production and as products sold by companies. Knowledge as such is normally not subject to tariff protection, and traditional models concerning protectionism are often of questionable relevance in the case of know-how intensive industries. Instead, in the light of certain empirical observations, real world governments have sometimes favoured their domestic companies at the cost of their foreign rivals in patent policies. Discrimination against foreigners can obtain in either patent legislation or administrative practice. It may be more difficult for a foreign company to get a patent, or alternatively the protection offered

by a patent may be weaker in the case of a foreign patentee. Aoki and Prusa (1993) presented a pioneering model to analyse the impact of discriminatory patent policies on firms' incentives to invest in R&D. In their relatively short paper they were not able to present a comprehensive analysis of all the potentially interesting issues related to discriminatory patent policies. Instead, they introduced some basic analytical tools that can be applied to this increasingly important topic.

In chapter 3 the basic analytical tools presented by Aoki and Prusa are used to analyse the topic further. We shall see that the welfare effects of discriminatory patent policies are entirely different from the effects of tariff protection imposed on foreign imports of physical goods.

The chapter presents two main findings. First, if two governments discriminate against each other's domestic firms, these policies may offset each other's effects, and the policies may be simultaneously both useless and harmless. At the firm level, the effects of being favoured in the home country and discriminated against in the foreign country may offset each other. Therefore discrimination may have no impact on a firms' R&D efforts. Secondly, it is demonstrated that unilateral discriminatory protection offered to a domestic company competing against a foreign rival can either encourage or discourage domestic R&D. If intensifying the R&D effort is useful mainly because additional expenditure increases the likely value of the invention, discriminatory protection would mainly reduce the most important risk related to the investment, namely the possibility that the rival would win the patent race. Obviously, reducing the risk would strengthen the incentive to invest in R&D. If, instead, additional R&D mainly increases the likelihood of gettting the patent but has a minor impact at most on the value of the patented invention, protectionist policies by the domestic government could be a good substitute for costly R&D and thus reduce it.

Though chapter 3 probably has few robust policy implications, it may be of interest because it questions conventional wisdom. Discrimination against foreigners in the case of intellectual property rights in oligopolistic industries has little to do with traditional tariff protection, and it may have entirely different effects. It is far from obvious that international agreements aimed at dismantling discriminatory patent legislation contribute to global welfare.

Chapter 4 analyses firms' incentives to patent their inventions. If firms produce more inventions, does the number of patent applications automatically increase? Because patent statistics are frequently used

as an indicator of technological progress in the previous literature, the relevance of several empirical papers depends on this issue.

It is relatively easy to find theoretical papers that analyse the incentives of firms to patent their inventions, for instance Horstman et al. (1985), Choi (1990), Saarenheimo (1994), Takalo (1996). But unfortunately this research tradition is surprisingly loosely related to the empirical literature. Chapter 4 hopefully narrows the gap between empirical and theoretical research.

A model inspired by previous empirical results is presented in Chapter 4. In the light of previous empirical findings, the correlation between R&D effort and patenting is much weaker at the firm level than at the industry level. Instead, patenting at the company level correlates rather strongly with rival R&D, even in the short term (eg Pakes and Griliches, 1984).[2] This might indicate that firms' patenting behaviour depends on their rivals' R&D efforts. How could this be explained? The model presents a potential explanation.

The basic idea of the model is quite simple. Firms can earn profits with their inventions if and only if the inventions are monopolized. Inventions can be monopolized either by patenting them or by keeping essential details secret. Because patent applications are public documents and thus a source of free information for other companies, and because patents do not offer perfect protection against imitators, secret inventions are more valuable than patented ones. On the other hand, by not patenting the firm runs the risk that a rival might invent the same technology. This risk is substantial if the rival invests heavily in R&D. But if the rival spends little on new technologies, there is no need to protect oneself by patenting. Thus firms' patenting policies reflect rival R&D intensity. Moreover, risk-averse firms are more willing to patent than are risk-neutral firms.

The model finds further empirical evidence in panel data estimations concerning the number of patent applications in different industries in Finland. When Finnish firms have intensified their R&D efforts, their foreign rivals have typically filed more patent applications in Finland. By contrast, there seems to be no immediate correlation between R&D efforts of Finnish firms and the number of patent applications filed by them. Assuming that foreign firms patent their research results in order to protect themselves against their Finnish rivals would explain these findings. Moreover, it is found that the number of domestic patent applications has been counter-cyclical

[2] See section 4.2.2 for more references concerning the covariation between patent counts and R&D efforts.

in the post-World War II era. If we assume that firms are decreasingly risk averse, this finding can be explained with the basic concepts of the model.

2 Developing an interbank payment system – Efficiency of public versus private investments

2.1 Background for the model

2.1.1 Purpose of the model

In many countries, the central bank is responsible for both price stability and the smooth functioning of payment systems. For instance, the Act on the Bank of Finland explicitly states (paragraph three) that the central bank shall contribute to developing the payment system. According to the Maastricht treaty, the European Central Bank is obliged to contribute to the smooth functioning of payment systems.

But if the system is maintained by the private sector with minimal central bank involvement, will there be a market failure? Because theorists have largely ignored this question, it is difficult to justify with solid economic arguments the existence of laws that oblige central banks to contribute to payment systems. This paper is a preliminary attempt to shed some light on this subject.

As we will show in this paper, the nature of optimal central bank involvement may depend on various factors. It turns out that optimal central bank policy may depend on whether payment services are a free service and on the market structure of the banking industry.

According to the model to be presented in the following, the market outcome is seriously distorted if payment services are offered to the public free of charge as a marketing tool. If customers' needs for services were small, the private sector would invest heavily, and vice versa. The incentives of the private sector are totally distorted in two extreme cases, namely if either the market is highly concentrated or the market shares are of equal size. The central bank should play an important role because private investment may be insufficient and because the central bank can affect the behaviour of the private sector.

If instead payment services are offered because of the fee revenue that can be earned, the situation is different. Banks can increase their income by improving the system when the number of interbank payments is large. Thus developing the system is profitable when there are a lot of interbank payments to be processed. This is a socially desirable incentive structure. Especially if banks' market shares are

equal, there will be no serious market failure, and involvement of the central bank is not as essential as with free payment services. In fact, excessive investments by the central bank might even worsen the allocative distortions that could emerge with the use of private resources.

As with many formal models, the one presented in the following sections can be interpreted in various ways. There are certainly strong analogies with payment systems and the telecommunications industry, and it might be possible to interpret the model as a description of competition between, say, two mobile phone operators. In fact, there are hardly any details in the following model that are absolutely inconsistent with the realities of the telecommunications industry. Thus the model might have some implications for other industries as well.

2.1.2 Central banks and interbank payments in the real world

International comparisons reveal that there are clear differences between countries in the degree of central bank involvement in payment systems. In Germany and the US, for example, a significant part of the payment system is virtually run by the central bank. In some other countries, such as the UK, the role of the central bank is limited to final settlements between a few major payment system agents.

The efforts of the central bank are often essential to the smooth transacting of interbank payments. In practice, it would be hardly possible to create a reliable and efficient interbank payment system without any involvement by the central bank. At least the final (net) settlement between banks is effected with central bank money.

There are at least three kinds of investments the central bank can make.

1) The central bank can make purely technical investments, such as renovations of its own computer software and hardware. For instance, it could offer various types of alternative settlement systems for payments made with central bank money, or arrange automatic queuing facilities to facilitate clearing in case of illiquidity. These improvements might reduce the number of errors and speed up the process. The central bank could also make direct contributions to systems that are owned and operated collectively by the government and the private sector.

2) The central bank can adapt its rules and practices so that the interbank payment system functions better. For instance, the frequency of net clearings could be increased, which would enable banks to arrange faster payment services. Or, to take another example, if the customers of a bank make significantly more payments than they receive, the bank might not be able to pay the sum of these payments to other clearing parties unless the central bank provides it with sufficient liquidity.

3) The central bank can arrange combinations of regulations and subsidies that lead to improved payment services. For instance, the central bank might require that at least certain payments are processed with real time gross settlement (which might be more burdensome for banks) and at the same time subsidise participating banks by offering free services.

2.1.3 The literature

There is a vast literature analysing the monetary policy function of central banks, and it is possible to identify different research areas within the field, such as central bank independence. By contrast, there are few studies that provide theoretical insights concerning a central bank's optimal payment systems policy. The functioning of different settlement systems is one of the few topics that has been analysed, but the focus here has been on differences between net and gross settlement systems rather than on market failures requiring governmental intervention.

There are several theoretical studies that analyse the specifics of the banking industry, as either an oligopoly or a monopoly. In these studies, the focus is on banks' role as financial intermediaries between savers and investors. One of the central topics has been the strengths and weaknesses of intermediated finance compared to direct market-based financing. These models are often based on asymmetric information. Banks supposedly can monitor their debtors better than savers can. The role of branch networks and ATMs in differentiation and oligopolistic competition has been another central topic.

Moreover, there are several studies that deal with payment systems. In most of these previous studies, the focus has been on the tradeoff between risks and cost efficiency in clearing and settlement. If banks do not coordinate their actions, they normally maintain suboptimal settlement balances, for instance on their central bank

accounts (see Angelini and Gianni, 1994; or Koponen and Soramäki, 1998).

Here, the aim is to analyse the topic from another point of view, namely by viewing the situation as a struggle for market share. Risks and costs related to different interbank net and gross settlement systems are ignored. One of the few previous theoretical studies on payment services as a competitive tool is that of McAndrews and Roberds (1997). They developed a duopoly model that describes cheque processing. A bank that can control the clearinghouse can either charge fees or delay cheque processing in order to adversely affect a competitor's ability to offer services.

The model presented in the following is characterized by network externalities. Several articles dealing with these effects have been written. The concept was introduced to economic theory in the mid-1980s, one of the first contributions being made by Katz and Shapiro (1985). In the presence of network externalities, the utility provided by goods is greater if the number of other consumers using the same product is large. The telephone and email are excellent examples of these kinds of effects; a telephone yields no consumer utility unless there are at least two telephones connected to the same network. One of the key issues that has been examined is the pricing policies of a monopoly company that owns the network.

Liebowitz and Margolis (1994) is a review of the literature in this field. They emphasise the difference between network externalities and mere network effects, and argue that pure network externalities are not commonplace. Network effects, by contrast, are frequently encountered. A network effect arises, for instance, because the availability of software improves as the number of consumers using a certain type of computer increases. Nevertheless, this effect is not an externality in the strict sense of the word, because the number of users does not directly enter the utility function of any consumer. The utility yielded by the computer and its software may not depend on the number of other consumers using the same standard.

The impact of network externalities in competition between firms has been the topic of several recent theoretical studies. Laffont et al. (1997) presented a model describing the competition between two telephone operators. The model has several strong analogies with the model presented below. There are two telecommunications operators in their model. Each customer is interested in communicating with other customers of the same operator and with customers of the other operator. In this model, nonlinear pricing and barriers to entry are analysed. Probably the most important analogy between this model and the model presented in this paper is the strategic motivation of the

bigger competitor to hinder connections between service providers. By contrast, Laffont, Rey and Tirole do not analyse investments aimed at developing the linkage between the networks. Neither is there a public body acting as the centre of the network in the Laffont-Rey-Tirole model.

2.2 Structure of the payment system in the model

The model describes a payment system consisting of providers of payment services and the customers who use the system for payment transactions among themselves.

There are two profit maximizing commercial banks that provide their customers with different banking services, including payment services. The payment services could be for giro, cheque or debit card payments, though the model probably best describes giro payments.

There are n customers, where n is very large ($n >> 0$). In the real world, it is likely that there would be millions of customers. Because the number of customers is very large, no single customer can affect market shares by his own decisions. Every customer has a client relationship with exactly one of the two banks. These customers have to make payments among themselves. The same customers are both payers and payees. Every customer has to make one payment to another customer in the same economy. The customer-consumers maximize their personal welfare, and the quality of payment services is one of the factors that affects their utility.

The payment system consists of the following subsystems.

1) Intrabank systems used in giro, cheque and debit card payments between customers of the same bank. Both banks have an intrabank system, the quality of which is exogenous.

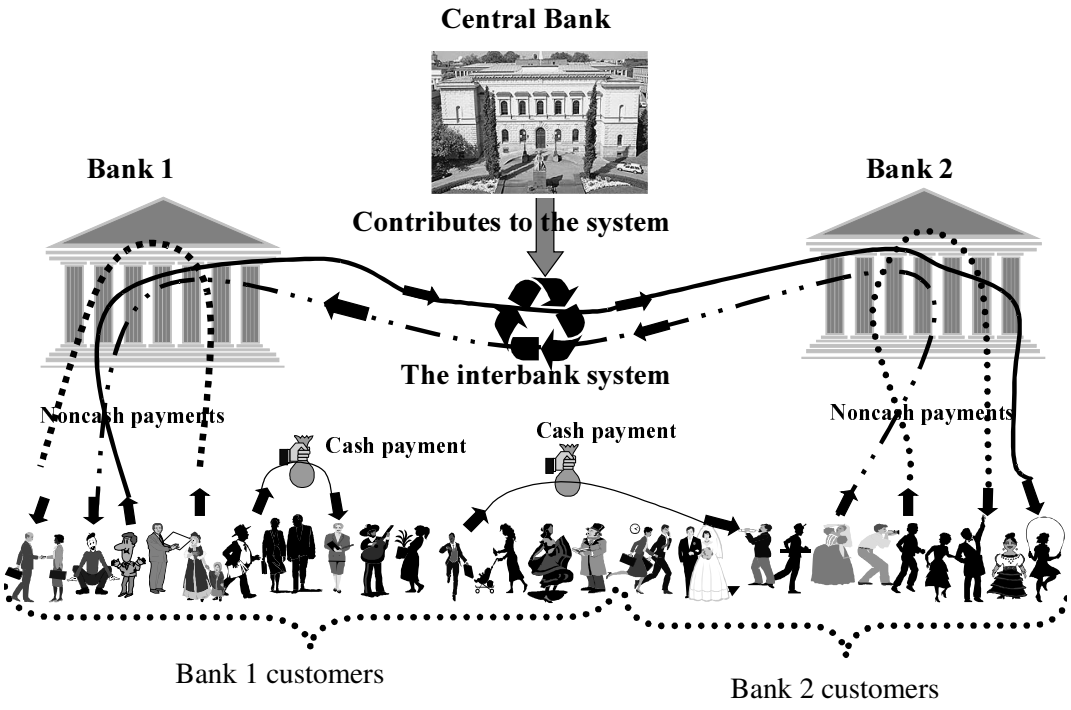
2) An interbank system for giro, cheque and debit card payments between customers of different banks. The interbank system is both developed and operated by the two commercial banks and the central bank. The quality of the interbank system is determined by the investments of the two private banks and the central bank.

3) Cash in circulation. Any consumer in the economy can make a cash payment to any other consumer in the economy. Customers do not need to use any banking services to make cash payments.

Due to lost interest reve-nue, the inconvenience of queuing at the ATM and the risk of theft, the utility of using cash as a means of payment is negative. The utility of a cash payment equals –w. This cost is evenly divided between payer and payee.

Figure 1 illustrates the role of the different agents in the payment system.

Figure 1. **Structure of the payment system**



There is no interbank payment centre owned and run by the banks collectively, and no collusion between banks. Exchange of payment information takes place either directly between the two banks or indirectly via the central bank. The (net) clearing services are arranged by the central bank.

It is assumed that there are no capacity constraints on the system. Once it is established and developed, any number of transactions can be processed. In this sense, the model describes the present situation rather than the past. Computer systems are normally expensive to establish, but the marginal cost of running them is negligible.

2.3 Basic version: Payments as a free service

In this version of the model, it is assumed that banks offer payment services free of charge. There could be a law that obliges them to offer free services, or the banks themselves could decide not to charge fees. It has also been proposed that banks might offer free or underpriced services because of tax incentives; customers prefer non-taxable benefits, such as access to free or subsidized services, as compared to taxable interest income (Tarkka, 1995).

2.3.1 Assumptions

2.3.1.1 Players' moves

The model is a full information game, in which agents can always correctly calculate each other's decisions and moves. Moves are made in the following order.

1) The central bank decides its own investment for development of the interbank payment system. Commercial banks can immediately observe the level of investment.

2) Commercial banks decide simultaneously on their investments in the development of the interbank payment system.

3) Customers observe the quality of the interbank system. Each customer chooses his bank. These choices are made independently and simultaneously without cooperation. When choosing his bank, a customer does not know to whom he will make his payment. All other customers are equally likely to be the payee.

4) Each customer observes to whom he will make his payment. The payees of different customers are determined independently.

5) Customers make and receive their payments.

2.3.1.2 Functioning of the interbank payment system

Once the payment system is developed, banks can use it at zero marginal cost. If there were a constant nonzero marginal cost per

transaction, the model might describe a situation where banks would be obliged to set the price at the marginal cost.

Developing the interbank payment system improves the quality of the service. Customers prefer a high quality system. The term 'quality' refers to the speed and reliability of the system. To concretize with extreme examples, if the system is of weak quality, it might take weeks for an employee to get his salary payment if his employer uses a different bank. When the electricity company tries to collect its receivables, it would have no problems with payments made by customers who use the same bank as the electricity company, but, if the customer uses a different bank, the company would never get the payment. If the quality of the interbank system is excellent, there is hardly any difference between making an interbank vs an intrabank payment.

The quality of the interbank payment system is a function of the development efforts by the three agents: payer's bank, payee's bank and the central bank.

The following notation is used.

Λ_1 = investment by bank 1,
Λ_2 = investment by bank 2 and
Λ_c = investment by the central bank.

The quality of an interbank payment is independent of whether it goes from bank 1 to bank 2, or *vice versa*. In both cases, the payment must go through the same chain, which consists of three systems. Quality is beneficial to both payer and payee. Customers prefer receiving payments through a highly developed system.

Interbank payments are possible (though of low quality) even if nothing has been invested in developing the system.

The number of customers (n) is very large. The market share of bank 1 being s, the probability that a payment will go to a customer of bank 1 is s, and the probability that it goes to a customer of bank 2 is (1–s).

2.3.1.3 Customers' preferences in a Hotelling duopoly

The total utility of a customer is determined by the quality of payment services and by customers' preferences as between the two banks.

Each consumer has to make a discrete choice between the two banks. Three factors are taken into account: First, the distance to the bank, secondly, a general preference parameter (G), and finally the

number of other customers who are going to use the same bank. Customers can correctly calculate others' choices and the resulting market shares.

Customers are risk neutral. Risk neutrality matters because the consumer does not know in advance to whom he will make his payments, nor from whom he will receive payments. The expected utility of consumer x is

$$W_x = \begin{cases} G + (2 - i_x) + (1 - s) \cdot a + s & \text{if he chooses bank 1} \\ -G + (i_x - 1) + s \cdot a + (1 - s) & \text{if he chooses bank 2} \end{cases}$$

where

– G is an exogenous preference parameter. The parameter is exogenous and common to all customers. If G > 0, bank 1 is preferred by most customers. If G < 0, most customers prefer bank 2. If G = 0, customers are, on average, indifferent between the two banks. This parameter does not reflect any scale or newtork effects, and its value is not affected by other customers' choices. To take a concrete example, there might be exogenous differences in the quality of customer services – or possibly there is a difference in the credit ratings of the two banks, which is relevant to customers' choices if deposits are not fully insured.[3]

– i_x is a customer-specific exogenous parameter, denoting the location of the customer.[4] The banks are located at the endpoints of the interval, bank 1 at point 1 and bank 2 at point 2. Getting service from a bank that is close to the customer provides the customer with higher utility. If $1 \leq i_x < 1½$, parameter i favours bank 1, if $1½ < i_x \leq 2$, parameter i favours bank 2, and if $i_x = 1½$, the parameter is neutral as between the two banks. Because the

[3] These assumptions imply that the average level of utility provided by the two banks must always equal 0. Some readers may find this assumption unrealistic. Fortunately, the assumption does not affect the results. Let R denote the average utility provided by banks' services. Then, by definition, if the utility provided by bank 1 equals R+G, then the utility provided by bank 2 must equal R–G. Thus R becomes an exogenous constant that enters the utility function of every customer, irrespective of banks' investments and irrespective of customers' choices between the two banks. Because the constant is entirely exogenous and has no effect on the workings of the model, nothing is lost by excluding it.

[4] The easiest interpretation of this parameter is that it describes the geographic distance, even though other interpretations are possible as well. For instance, bank 1 could use Swedish as its customer service language and bank 2 Finnish.

common parameter G may differ from zero, i_x < 1½ does not necessarily imply that consumer x would prefer bank 1.

– s = the endogenously determined market share of bank 1.

– a is the quality of the interbank payment system. If the value is high, making and receiving interbank payments is easy, efficient and reliable. If its value is low, execution of interbank payments might be slow and unreliable. The quality is always lower than the quality of intrabank payments. The total utility provided by an interbank payment equals its quality, a, which is evenly divided between payer and payee. The expected value of the number of interbank payments to be paid by a customer of bank 1 is (1–s), and the expected value of the number of interbank payments to be *received* is also (1–s). Therefore, the expected utility of interbank payments is 2·(a/2)·(1–s) = a(1–s). Analogically, the expected utility provided by interbank payments for a bank 2 customer equals 2·(a/2)·s = s·a. a must be positive, but it cannot be greater than +1.

– If the customer has chosen bank 1, the probability that a payment will go to another customer of the same bank equals its market share (s). Analogically, the expected value of the number of intrabank payments to be *received* is s for a bank 1 customer. The utility of being involved in one intrabank transaction is assumed to equal ½. The customer benefits from intrabank payments as both payer and payee, implying that the expected utility provided by intrabank transactions is 2·½·s = s for a bank 1 customer. The expected utility provided by intrabank payments for a bank 2 customer is determined in an analogous way to be 2·½·(1–s) = (1–s).

With the exception of the preference parameter i_x, all parameters are common to all customers.

There is a pure network externality in the model. Using the same bank as the majority of customers provides the consumer with utility. This network effect is a direct externality and is not caused by the effects of other customers' choices on any prices.

The model describes a giro transfer system rather than a cheque-based system. In a cheque system, it is far from obvious why customers would benefit if the interbank system were improved. In fact, they might prefer a slow and unreliable system. At least the payer would gain marginal interest income with lengthy delays between

moment of ordering a payment and moment of debitting of his account. If the system were so unreliable that a significant part of the cheques were lost in interbank clearing, with no debitting of the payer's account, customers would be even better off. (However, if it were commonplace to debit the wrong account, customers would probably prefer a more reliable system.)

2.3.1.4 Quality of the interbank payment system

There are two kinds of noncash payments.

1) Intrabank payments between two customers of the same bank. The quality of an intrabank payment is exogenous and equal to +1.

2) Interbank payments. The quality of an interbank payment is denoted a.

The value of the parameter a is a function of the investments made by the two banks and the central bank. If the value of a is high, payments are processed fast and reliably. If a is close to zero, interbank payments are slow and unreliable.

Both private banks can affect the quality of the system by investing in it. In addition, investments by the central bank affect the quality. These investments have a declining marginal impact on the quality. With zero investment by any of the three agents, the marginal impact of investment on the quality is infinite. The investments are allowed to have different interaction effects on the quality.

The following assumptions characterize the 'a' function:

1) It is possible to make payments between the two banks even if nothing has been invested in the system, ie for any level of investment, $a > 0$.
2) Investing in the system always improves its quality, although this improvement is subject to diminishing returns: $\partial a/\partial \Lambda_1 > 0$, $\partial a/\partial \Lambda_2 > 0$, $\partial a/\partial \Lambda_c > 0$, $\partial^2 a/\partial \Lambda_1^2 < 0, \partial^2 a/\partial \Lambda_2^2 < 0, \partial^2 a/\partial \Lambda_c^2 < 0$.
3) If an agent has not previously invested in the system, any investment would have an enormous impact on the quality of the system: If $\Lambda_1 = 0$, then $\partial a/\partial \Lambda_1 = \infty$; if $\Lambda_2 = 0$, then $\partial a/\partial \Lambda_2 = \infty$; if $\Lambda_c = 0$, then $\partial a/\partial \Lambda_c = \infty$.

4) The cross partial derivatives may be positive, negative or zero, but they are always finite: $|\partial^2 a/\partial\Lambda_1\partial\Lambda_c| \neq \infty$, $|\partial^2 a/\partial\Lambda_2\partial\Lambda_c| \neq \infty$, $|\partial^2 a/\partial\Lambda_1\partial\Lambda_2| \neq \infty$.
5) The lowest possible value for any investment variable Λ is 0. No agent can make a negative investment.
6) The 'a' function depends in a similar way on Λ_1 and Λ_2. If $\Lambda_1 = A$ and $\Lambda_2 = B$, the value of 'a' is equal to the value of 'a' obtained when $\Lambda_2 = A$ and $\Lambda_1 = B$. As to derivatives, the value of $\partial a/\partial L_z$ (or $\partial^2 a/\partial L_c\partial L_z$ or $\partial^2 a/\partial\Lambda_z^2$) depends on the value of L_z, not on whether z = 1 or 2.[5]
7) A payment from one bank to another can never be of higher quality than an intrabank payment, which is not transferred between banks. Even in the best case, the quality of interbank payments is at least marginally lower than the quality of intrabank payments, ie $a < 1$.

In the real world, banks may prefer slower interbank payments to faster ones. With slower processing of payments, banks can earn additional interest income on the float; if the account of the payer is debited several days before the account of the payee is credited, total interest payments to deposit customers are lower. Thus, system development is not necessarily a technological effort. One possible interpretation of the model is that the 'investment expenditure' consists partly of the interest loss caused by the decision to speed up the payment process with existing technological facilities.

2.3.1.5 Banks' revenues and profits

In addition to payment services, banks offer loan and deposit services as well, even though these are not explicitly modelled. In addition, each customer causes the bank some costs. Computer systems and the physical retail service network are more expensive to maintain if the bank has a lot of customers. The parameter α describes the exogenous net income per customer that a bank earns in collecting deposits, granting loans and maintaining the necessary infrastructure. The parameter is common to all customers and both banks. Therefore, the total net income of a bank equals α times the number of customers.

[5] This does not necessarily imply that $a = a\{(\Lambda_1 + \Lambda_2), L_c)\}$.

The profit of bank 1, when the cost of payment system development is not taken into account, is

$$\Pi_1 = n \cdot s \cdot \alpha$$

and the profit of bank 2 is

$$\Pi_2 = n \cdot (1 - s) \cdot \alpha$$

The sum of the banks' profits is always na. Thus, the struggle for market share is a zero sum game between duopolists. A bank can increase its profits only at the cost of its rival.

2.3.2 Solving the model

2.3.2.1 Banks' market shares

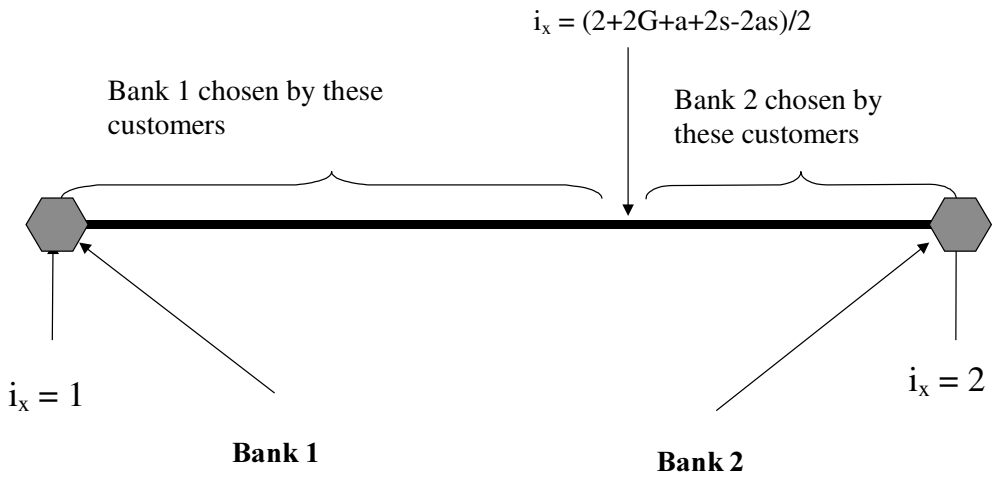
The utility of consumer x is determined according to the function described in section 2.3.1.3.

The customer chooses bank 1 if

$$\begin{aligned} & G + (2 - i_x) + (1 - s) \cdot a + s > -G + (i_x - 1) + s \cdot a + (1 - s) \\ & \Leftrightarrow i_x < (2 + 2 \cdot G + a + 2 \cdot s - 2 \cdot a \cdot s)/2. \end{aligned} \qquad (2.3.i)$$

The market share of bank 1 is determined by the number of customers for whom the condition is valid.

Figure 2. **How customers choose a bank**



The density function for customers is extremely simple, ie the constant +n. Hence, because n is a very large integer, the market share is almost exactly

$$s = \left\{ \int_1^c n \, di \right\} / n.$$

where c is the point where the condition for i_x (2.3.i) is no longer valid, ie $c = (2+2 \cdot G+a+2 \cdot s-2 \cdot a \cdot s)/2$.

The market share is given by

$$s = \frac{(2G + a)}{(2a)} = ½ + \frac{G}{a}. \quad (2.3.ii)$$

Thus bank 2's market share is $1-s = 1-(2G+a)/(2a) = (-2G+a)/(2a) = ½-G/a$.

The market share of bank 1 is between 0 and 100 % if $|2G| \leq a$. If $-a/2 < G < 0$, bank 1's market share is between 0 and 50 %. And if $0 < G < a/2$, the bank's market share is between 50 % and 100 %.

If $G < -a/2$ ($G > a/2$), the formula for s would predict bank 1's market share to be less than 0 (greater than 1). In practice, in such cases, the market shares would be 100 % and 0 %, which would fall outside the duopoly case. The monopoly situation would be a completely meaningful case, but interbank payments would not take place in such a banking industry, and the following analysis would not apply. Hence, in the following, it is assumed that each bank has a

positive market share, even though a market share may be close to zero.

As to real life interpretations, cases where G is close to either –a/2 or a/2, are not very interesting. Almost all countries have certain government regulations concerning minimum capitalization of credit institutions. In the EU, for instance, a bank with equity capital of less than ECU 5 million is not allowed to enter the market. Thus the smallest bank that can legally exist is a local institution with a moderate market share in its home area.

The situation is fundamentally symmetric, the only difference between the two banks being the eventually nonzero value of G. Therefore, all the following results are equally valid for both banks. To simplify the notation, the analysis is in most cases presented only for bank 1.

2.3.2.2 Banks' investments in the payment system

From the point of view of a bank, the quality of the payment system is relevant to profits because it affects market shares. The impact of payment system development on the market share of the bank can be calculated using the formula (2.3.ii)

$$\frac{ds}{da} = \left(\frac{-G}{a^2} \right)$$

If the two banks are equally popular (G = 0), then the quality of the payment system is of no relevance to market shares (ds/da = 0). Hence, the more popular the bank, the less system development helps to increase market share

$$\frac{d^2 s}{da\,dG} = \frac{-1}{a^2} < 0.$$

If the market share is greater than 50 %, then developing the payment system causes a loss of customers to the rival. A well functioning interbank system would improve the payment services received by all the customers, but the effect would be even stronger for customers of the small bank because most of their payments are interbank payments. Therefore, the improvement would strengthen the competitive position of the smaller bank. If the interbank payment system does not function smoothly, most customers will find it

advantageous to use the more popular bank. A customer of a small bank would be nearly isolated if the interbank payment system did not function properly, which would cause him a substantial amount of disutility.

Result 2.3.a;
If market shares are equal, neither of the banks will invest in developing the system. If the market shares are not equal the smaller bank invests in developing the system but the larger does not.

Proof

$$\frac{d\Pi_1}{da} = n\alpha\frac{ds}{da} = -n\alpha\cdot\left(\frac{G}{a^2}\right).$$

If $G \geq 0$, the bank has no incentive to develop the system because

$$\frac{d\Pi_1}{da} = -n\alpha\cdot\left(\frac{G}{a^2}\right) \leq 0.$$

If $G < 0$, then $d\Pi_1/da = -n\alpha\cdot(G/a^2) > 0 \Rightarrow$ bank 1 does have an incentive to invest in the system.

When the investment by the bank 1 (Λ_1) is zero, then, by assumption, $\partial a/\partial\Lambda_1 = \infty$.

If $d\Pi_1/da > 0$ and $\partial a/\partial\Lambda_1 = \infty$, then it is optimal for bank 1 to invest in developing the system.

QED

Unsurprisingly, the willingness to invest in developing the system is a decreasing function of the popularity parameter, G.

Result 2.3.b
If bank 1 is less popular than bank 2 (which will be the case when $G < 0$), then an increase in the popularity of bank 1 (G) will cause it to invest less in developing the system.

Proof

The bank maximises its profits ($\Pi_1 - \Lambda_1$) according to the first order condition $\partial\Pi_1/\partial\Lambda_1 - 1 = 0$.

Implicit differentiation gives $d\Lambda_1 / dG =$
$-\left\{\partial^2\Pi_1 / \partial\Lambda_1\partial G\right\}/\left\{\partial^2\Pi_1 / \partial\Lambda_1^2\right\}$

The second order condition for profit maximization implies $\partial^2\Pi_1 / \partial\Lambda_1^2 < 0$.

One can write $\partial\Pi_1/\partial\Lambda_1 = n\cdot\alpha\cdot(ds/da)\cdot(\partial a/\partial\Lambda_1)$ and $d^2\Pi_1/d\Lambda_1\, dG = n\cdot\alpha\cdot(d^2s/dadG)\cdot(\partial a/\partial\Lambda_1)$.

Because
$d^2s/dadG = -1/a^2 < 0$, $n\cdot\alpha > 0$, $da/d\Lambda_1 > 0$ and $d^2\Pi_1/d\Lambda_1\, dG < 0$, it follows that $d\Lambda_1/dG < 0$.

QED

2.3.2.3 Banks' actual investments vs socially optimal investments

As concluded above, only the smaller bank invests in system development. And if neither of the banks is smaller than the other, there is no private investment at all. As a rule, this is not socially optimal. Because the marginal impact of investing in the system is extremely high when investment is close to zero, both of the banks should invest equally heavily, if the system is to be developed at all. This would be the most cost-efficient way to reach any given level of *a*. Therefore, there is an obvious market failure. Moreover, in most cases, the investment by the smaller bank differs from the socially optimal level.

The struggle for the net interest income ($n\cdot\alpha$) is a zero-sum game between the two banks. Therefore, when one analyses the social welfare effects of payment system development, one can focus entirely on the utility consumers get from using the system.

As to a bank 1 customer, his utility equals

$$W_x = G + (2 - i_x) + s + (1 - s)\cdot a$$

and the utility of a bank 2 customer equals

$$W_x = -G + (i_x - 1) + s \cdot a + (1 - s).$$

Total welfare for the economy equals

$$\psi = \sum_{i=1}^{z} W_i + \sum_{i=z+1}^{n} W_i + \Pi_1 + \Pi_2 - \Lambda_1 - \Lambda_2$$

where z = n·s = the number of the last customer to use bank 1.

U_z (z = 1, 2) denotes the utility yielded by payment services to a customer of bank z. $U_1 = s + (1 - s)\cdot a$, and $U_2 = s\cdot a + (1 - s)$. Thus, if the customer uses the bank 1, then $W_x = G + (2 - i_x) + U_1$, and if the customer uses the bank 2, then $W_x = -G + (i_x - 1) + U2$.

The subutility function U_z does not depend on the individual preference parameter i; thus it does not vary as between individuals. The derivative dU_z/da equals the whole impact of a on the welfare of a bank z customer. Because the values of the preference parameters (G and i) do not have any impact on the effect, all bank z customers have an equal value for dU_z/da. If customer i uses bank 1, then $dW_i/da = dU_1/da$.

The impact of payment system development on social welfare (ψ) is

$$\frac{d\psi}{da} = n \cdot \left[s \cdot \left\{ \frac{dU_1}{da} \right\} + (1 - s) \cdot \left\{ \frac{dU_2}{da} \right\} \right]. \qquad (2.3.iii)$$

Result 2.3.c
The impact of the quality of the payment system on social welfare is maximal when G = 0

Proof

According to (2.3.iii)

$$\begin{aligned} \frac{d\psi}{da} &= n \cdot \left[s \cdot \left\{ \frac{dU_1}{da} \right\} + (1 - s) \cdot \left\{ \frac{dU_2}{da} \right\} \right] \\ &= n \cdot \left[s \cdot \left\{ \frac{½ - G}{a^2} \right\} + (1 - s) \cdot \left\{ \frac{½ + G}{a^2} \right\} \right] \\ &= n \cdot \left(\frac{½ - 2G^2}{a^3} \right). \end{aligned}$$

When this is differentiated with respect to G, one gets $-n4G/(a^3)$.
This equals 0 when G = 0.
Because $d^3\psi/da\ dG^2 = -n4/(a^3) < 0$, this is the maximum of $d\psi/da$.

QED

This result is easy to understand intuitively. If customers are evenly distributed between the two banks, the number of interbank payments is maximal. Therefore, improving the system would be highly useful. However, no private investment is made because neither of the two banks can increase its profits by such investments. As concluded above, in this special case, system development has no impact on market shares.

Result 2.3.d
If $|G| \geq a^{(3/2)}/2$, then the socially optimal investment in the system is zero.

Proof

$d\psi/da = n[½ - 2G^2/(a^3)]$. If $|G| > |a^{(3/2)}/2|$ then $d\psi/da < 0$, and therefore no investment would yield any social benefits.

QED

This result may sound rather counter-intuitive. If improved payment systems make interbank transactions fast and reliable, how could such a development be undesirable? The answer is related *to network externalities caused by consumer choice*. Suppose bank 1 has a very small market share. There is a customer who is indifferent between the small bank 1 and the large bank 2. The consumer might choose bank 2 at random. Then, a marginal improvement in interbank payments takes place; using bank 1 becomes slightly more attractive for the consumer because the problems created by the frequent need to make interbank transactions are marginally alleviated. The customer shifts to the small bank 1. This has a very small impact on personal utility, because the customer is nearly indifferent between the two banks. For other customers, this choice is more significant. This decision has a positive externality for all bank 1 customers and a negative externality for all bank 2 customers. The former group can exchange payments

with this particular customer with less difficulties than before, but the latter group suffers from a comparable negative externality. If the market share of bank 1 is very small, the resulting negative externality is strong enough to more than offset the benefits of improved interbank payment services. In such a case, the overwhelming majority of customers would find it more difficult to exchange payments with the customer who decided to choose the smaller bank. In principle, the assumptions of this model imply that a banking monopoly would be ideal for the payment system, but improving the payment system strengthens the relative position of the smaller bank.

This effect is a good example of what Liebowitz and Margolis (1994) classify as a direct network externality. The impact of consumer choice on other consumers is not channelled through the price mechanism. Instead, the choice itself has a direct impact on others' welfare.

As a rule, the investment by bank 1 is not socially optimal. However, investment by bank 1 is at its optimal level in two different cases. First, if bank 1's market share is large enough, increasing the market share of the minor bank 2 by improving the system would be socially undesirable because of its adverse impact on the payment system. Moreover, it would be unprofitable for bank 1 itself. Secondly, investment by bank 1 is at its optimal level at a particular point where the bank has a market share between 0 and 50 %.

Result 2.3.e
There is no market failure in by bank 1's investment in two cases, namely

1) **with one particular negative value of G, namely** $\mathbf{G = (a/4)\{\alpha - \sqrt{[4a+\alpha^2]}\}}$
2) **when** $\mathbf{G > a^{(3/2)}/2}$

Proof

There is no market failure if private and social benefits are equal, ie if

$$n \cdot a\left(\frac{-G}{a^2}\right) = n\left[\frac{½ - 2G^2}{\left(a^3\right)}\right].$$

This equation holds if $G = (a/4)\{\alpha \pm \sqrt{[4a+\alpha^2]}\}$. If $G > 0$, profit maximizing investment would be negative $\Rightarrow$ The eventual candidate is $G = (a/4)\{\alpha - \sqrt{[4a+\alpha^2]}\}$.

The market share of bank 1 is between 0 and 50 % if $(-a/2) < G < 0$.

With the value of G as given above, this condition is

$$-\left(\frac{a}{2}\right) < \left(\frac{a}{4}\right)\left\{\alpha - \sqrt{4a+\alpha^2}\right\}$$

or

$$2 + \alpha > \sqrt{\left[4a + \alpha^2\right]},$$

which simplifies to

$$4 + 4\alpha + \alpha^2 > 4a + \alpha^2;$$

Because $a < 1$ and $\alpha > 0$, this holds with certainty.

Because $a/4 > 0$ and $0 < \alpha < \sqrt{[4a+\alpha^2]}$, $(a/4)\{\alpha - \sqrt{[4a+\alpha^2]}\} < 0 \Rightarrow G < 0$. Therefore there is one value of G that implies a market share s ($0 < s < ½$) where bank 1 invests the socially optimal amount.

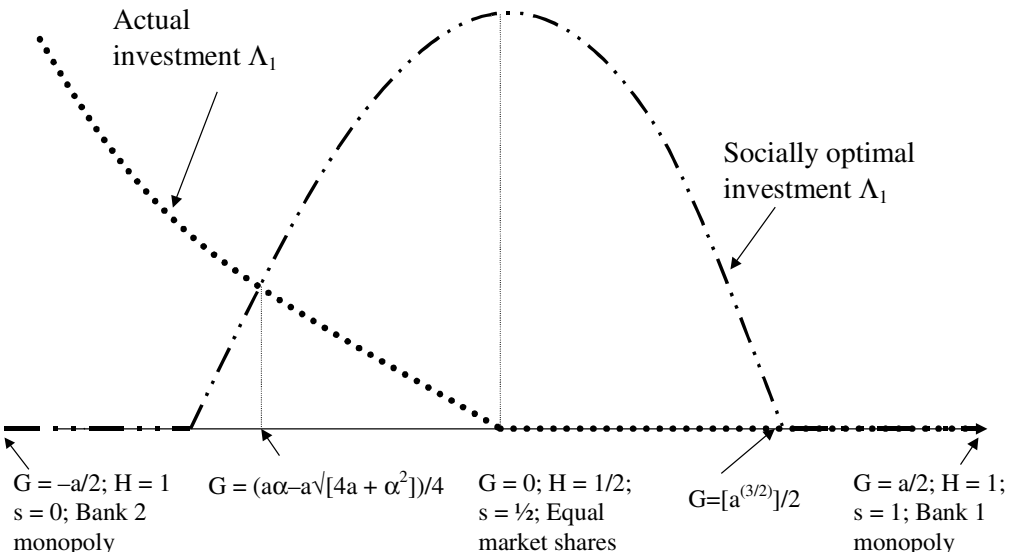
Moreover, there is another case where there is no market failure. The marginal impact of 'a' on bank 1's profits is negative if it has a greater-than-50 % market share. In this case, the bank does not invest in developing the system. This is socially optimal if $G > a^{(3/2)}/2 \Rightarrow d\psi/da < 0$ (result 2.3.d).

QED

Figure 3 may shed some light on market failure in respect of bank 1's investments. The socially optimal investment reaches its maximum when market shares are equal (result 2.3.c) and, if $|G| \geq a^{(3/2)}/2$, the socially optimal in-vestment is zero. If bank 1 has a small market share [$G < (a/4)\cdot(\alpha - \sqrt{[4a + \alpha^2]})$], it invests more than the socially optimal amount (results 2.3.a, 2.3.d and 2.3.e). And if its market share

is greater than 50 %, it invests nothing (result 2.3.a). The Herfindahl index (H) measures the degree of market concentration; in this case, it is defined as $s^2 + (1 - s)^2$.

Figure 3. **Actual investment by bank 1 and the socially optimal investment as a function of the market share (s), the parameter G and the Herfindahl-index (H)**



However, even though bank 1's investment may sometimes be at its socially optimal level, it is not possible to find examples where both banks would invest optimally, because it is never the case that they would both invest. Thus there is no market structure (value of G) that would lead to an efficient allocation of resources.

2.3.2.4 Optimal central bank involvement

The role of investments by the central bank will be analysed in this section. The central bank can affect social welfare via its payment system investments in two different ways:

1: Directly, through the impact of investment on the quality of the system.
2: Indirectly, in that investment by the central bank normally affects private sector investment.

In many cases, central bank investment reduces private investment. This is possible even when central bank investments strengthen the impact of private investments on the quality of payment systems (when $d^2a/d\Lambda_c d\Lambda_1 > 0$).

Result 2.3.f
Let the market share of bank 1 be less than 50 %. A sufficient condition for increasing central bank investment to reduce investment by bank 1 is that either (or both) of the following two conditions holds:

1) $\mathbf{(\partial^2 a/\partial\Lambda_1\partial\Lambda_c\} \leq 0.}$
2) $\mathbf{\Lambda_c = 0}$

Proof: See appendix 1.

It is fairly obvious that when central bank investment makes private investment inefficient ($\partial^2 a/\partial\Lambda_1\partial\Lambda_c \leq 0$), increasing public investment discourages private investment. This could be the case if, for instance, there is an investment that can be made by either the public or the private sector, but for both to do it would be wasteful.

It may be somewhat more difficult to understand why the result might always apply when central bank investment is close to zero ($\Lambda_c = 0$). The reason is simple: Central bank investments may discourage a private bank from investing in the system simply by affecting the quality of interbank payments (a). If the payment system already functions properly, private bank 1 does not have to invest in the system itself.

The total impact of central bank payment system development on social welfare equals the difference between the utility provided by a marginal improvement in payments and the increase in the costs of developing the system. Mathematically, the social welfare impact of a marginal increase in central bank investment is

$$\frac{d\psi}{d\Lambda_c} = \left(\frac{\partial\psi}{\partial a}\right)\cdot\left[\left(\frac{\partial a}{\partial\Lambda_c}\right) + \left(\frac{\partial a}{\partial\Lambda_z}\right)\cdot\left(\frac{\partial\Lambda_z}{\partial\Lambda_c}\right)\right] - \left(\frac{\partial\Lambda_z}{\partial\Lambda_c}\right) - 1 \qquad (2.3.iv)$$

where

– $(\partial\psi/\partial a)$ equals the improvement or deterioration in welfare due to the improvement in the payment system

– $(\partial a/\partial \Lambda_c)$ equals the direct impact of central bank investment on the quality of the payment system
– $(\partial a/\partial \Lambda_z)\cdot(d\Lambda_z/d\Lambda_c)]$ equals the indirect impact via the impact of central bank investment on private investment. z = 1 if bank 1 has a market share of less than 50 %; z = 2 if bank 2 has a market share of less than 50 %. (The larger bank does not invest.)
– $(d\Lambda_z/d\Lambda_c)$ equals the impact of central bank investment on private investment expenditure
– 1 = the marginal cost of investment by the central bank.

Because central bank investment affects the choices of the private sector, it is not surprising that in the maximization of social welfare the indirect effects of investments should be taken into account as well.

If it is certain that central bank investment reduces private investment, the optimal policy must clearly be the following.

Result 2.3.g
If the smaller bank invests less (more) than the socially optimal amount in payment system development, the central bank should restrict (increase) its investment in the system in order to encourage (discourage) private investment, at least if
$(d^2a/d\Lambda_1 d\Lambda_c\} \leq 0$.

Proof: See appendix 2.

On the other hand, it is much more difficult to draw robust conclusions concerning optimal central bank investment when $(\partial^2 a/\partial \Lambda_1 \partial \Lambda_c) > 0$. Even if the level of central bank investment were very low, result 2.3.g would still be valid, because central bank investment would reduce private investment. However, with higher levels of central bank investment, the result would be the reverse.

If market shares are evenly distributed (G = 0), the central bank cannot do much to affect private sector investment. In this case, the banks would neither increase nor decrease their investment in response to central bank investment, because there would be no private investment anyway. There are no private reactions to be taken into account by the central bank.

It may be surprising that central bank investments in the system may be useful even when there is a very small private bank that has almost no customers but is about to get a handful of them. An improvement in the system causes a direct reduction in social utility

because it affects market shares in a non-desirable way but, on the other hand, central bank investment would have beneficial indirect effects. A private bank with a very small market share invests excessively in the payment system, and a feasible way to reduce private investment is to have public investment. At least a very small amount of public investment could be justified in a highly concentrated market.

Result 2.3.h
It is optimal for the central bank to invest at least something in the system with any value of G, at least if $\alpha = 1$ and $\partial^2 a / \partial \Lambda_1^2$ is close to zero.[6]

Proof: See appendix 3.

The economic intuition behind this result is as follows. When the fixed income per customer (α) has a suitable value, the smaller bank has moderate incentives to attract customers if it is possible at a reasonable cost. A marginally positive investment by the central bank can reduce the total investment rather efficiently, because the incentive of bank 1 to increase its market share is fairly weak. If instead, private investment were close to zero because of a very low net income per customer, it would not be worthwhile to try to affect private investment. The impact of public investment on private investment is even stronger if the optimal amount of private investment is easily affected by changes in parameter values. This is the case when $d^2 a / d\Lambda_1^2$ is close to zero.

Optimal central bank investment policies when it is always optimal to invest something can be clarified with figure 4. In figure 4, it is assumed that central bank investment always reduces private investment and that the investment of a private bank responds strongly to changes in investment by the central bank. The optimal central bank investment is compared to a hypothetical case where private investment does not depend at all on central bank investment. (This

[6] Because $\frac{d^2\Pi_1}{da^2} < 0,\ \frac{\partial^2 a}{\partial \Lambda_1^2} = 0,\ \partial^2 a / \partial \Lambda_1^2 = 0,$ implies $\frac{\partial^2 \Pi_1}{\partial \Lambda_1^2} < 0,$ which is a necessary condition for the existence of an optimal finite amount of investment.

would be the case, for instance, if the private bank were unable to observe the investment by the central bank.) Now let us review how the central bank should alter its investment with different values of G.

If bank 1 has a very small market share ($G \approx -a/2$), it invests excessively in payment system development, even though system improvement yields negative social benefits and the investment generates costs. Even though improving the system has a negative net contribution to social welfare, a small investment by the central bank would be justified because it would strongly reduce private expenditure in the system.

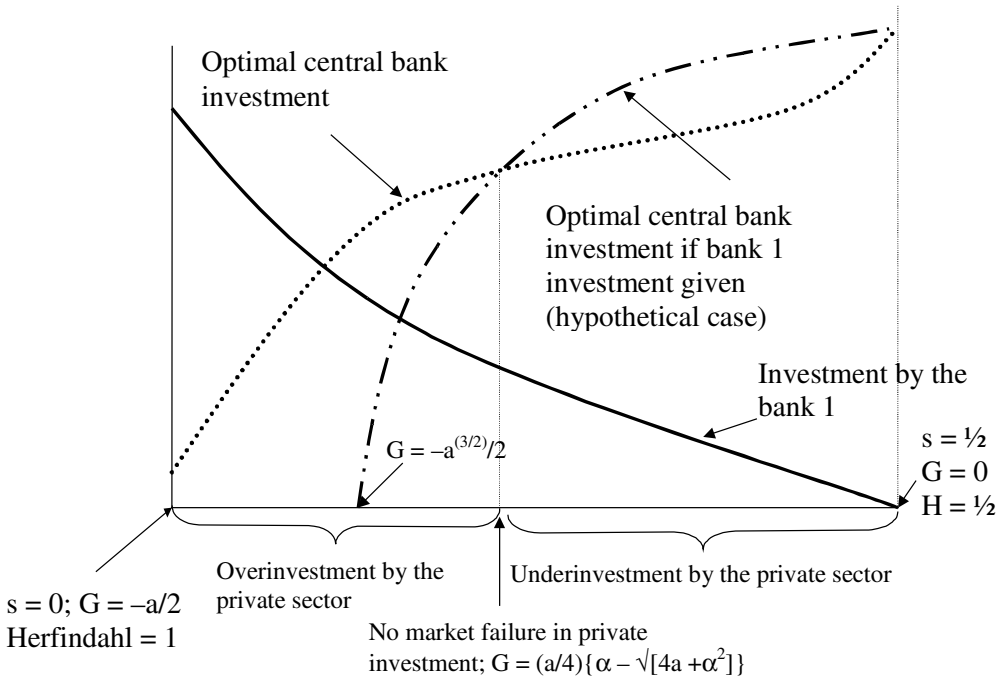
If G is somewhat higher than $-a^{(3/2)}/2$, then an exogenous improvement in the payment system would improve social welfare. Therefore, at least a small investment by the central bank would be socially optimal irrespective of reactions by the private bank. However, the social benefits of a better interbank payment system (higher a) are still lower than the marginal cost of improving the system via private investments. Therefore, the central bank should still try to restrict private investment by intensifying its own investment.

When bank 1's market share increases further and reaches the value implied by ($G = (a/4)\cdot\{\alpha - \sqrt{[4a + \alpha^2]}\}$), there will be enough interbank payments to make the social benefits equal the private cost of investment, implying that the central bank has no incentive to try to affect private investment. Thus the reactions of the private bank do not affect optimal central bank investment.

Beyond this point, central bank objectives change. Social benefits exceed private benefits, and bank 1 invests less than the socially optimal amount in developing the system. Therefore, the central bank should encourage private investment by restricting its own investment. As can be seen in figure 4, whenever the private bank 1 underinvests, optimal central bank investment exceeds the level that would be desirable if central bank investment did not affect private investment.

And finally, when $G = 0$, the private bank no longer invests, and the central bank cannot affect its behaviour. At this point, the number of interbank payments is maximal, and the only agent in the economy that invests in the payment system is the central bank. Therefore, it should invest substantially.

Figure 4. **Optimal central bank investment**
In this example optimal central bank investment is always greater than zero



Needless to say, it is possible and even simpler to construct examples where at low values of G, optimal central bank investment is zero. In such cases, the optimal central bank investment curve would cross the horizontal axis at a point between –a/2 and $-a^{(3/2)}/2$.

2.4 Bertrand competition in payment services

2.4.1 Assumptions

This version of the model differs from the basic model in that it is no longer assumed *a priori* that payment services are offered free of charge. Banks as providers of payment services are now allowed to charge fees, if they prefer to do so.

Moreover, it is no longer assumed that the fixed net income per customer is always positive. The fixed cost caused by one customer may or may not be higher than the net interest revenue per customer.

When banks have invested in developing the system, they must decide on their pricing policy. Banks decide all prices simultaneously in a Bertrand-type competition. The price charged by bank z for one interbank payment is denoted p_z (z = 1 or 2). Intrabank payments can also be priced, and the price of one intrabank payment is denoted b_z. Because of spatial differentiation, the two banks are not perfect substitutes, and they have a certain amount of market power. Bertrand competition does not lead to zero profits.

A bank may charge different prices for interbank and intrabank payments ($p_z \neq b_z$), if it prefers to do so. However, there is a marginal administrative cost (ε) that the bank must pay if it charges different prices for these two types of payments. This cost consists of the minor expenditures of adapting the information systems and informing employees concerning the two different prices. Compared to other expenses and revenues, this cost component is negligible. But if the bank is otherwise indifferent between a pricing policy characterized by $b_z = p_z$ and some other set of prices, then the bank prefers charging the same price for both interbank and intrabank payments. It is assumed that banks cannot charge anything for receiving payments.

Banks cannot practice price discrimination, possibly because they cannot observe the exact location of different consumers on the interval [1,2]. Payment service fees charged by banks cannot be negative; it is not possible for banks to pay their customers for using the service. But it is possible not to charge any fees.

There is no demand for interbank payment services if their price exceeds the reservation price, which equals $(a + \omega)/2$, where ω is the sum of the disutility of using cash suffered by payer and payee. For intrabank payments, the reservation price is $(1 + \omega)/2$.

There is no price elasticity of demand for payment services, provided the price is lower than the reservation price. A bank's service fee is such an insignificant cost that it cannot affect transactions in the real economy. However, the fee is relevant to the choice of payment medium. When customers have observed the fees, they decide which bank to use. At this stage, banks can no longer alter their prices. The central bank does not charge any fees.

If no customers use cash as a means of payment, the profit of bank 1 is

$$n \cdot s \cdot \alpha + n \cdot s(1-s) \cdot p_1 + n \cdot s^2 b_1$$

where n is the total number of customers in the economy (n>>0) and s = the market share of bank 1; p_1 = the fee charged for an interbank

payment by bank 1, and b_1 is the fee charged for an intrabank payment by bank 1.

In an analogous way, bank 2's profit is

$$n \cdot (1-s) \cdot \alpha + n \cdot (1-s) \cdot s \cdot p_2 + (1-s)^2 n \cdot b_2 .$$

The difference between the net interest income and fixed costs per customer (α) is still treated as exogenous, but as we shall see later (footnote 6), treating it as an endogenous variable would not affect the results significantly.

2.4.2 The Bertrand competition outcome

2.4.2.1 Banks' market shares

Now we will see how the market shares of banks are determined when the customers of both banks prefer interbank giro transfers to cash payments because prices are below their reservation levels. If both banks charge a price that is lower than the reservation price, customers prefer banks' payment system and the market shares are determined as follows:

Customer x chooses bank 1 if

$$G + (2 - i_x) + (1-s) \cdot (a - p_1) + s \cdot (1 - b_1) >$$
$$-G + (i_x - 1) + s \cdot (a - p_2) + (1-s)(1 - b_2)$$

which implies

$$i_x < \frac{(2 + a + b_2 + 2G - p_1 + 2s - 2as - b_1 s - b_2 s + p_1 s + p_2 s)}{2}.$$

Bank 1's market share is determined by the number of customers for whom the condition presented above is valid. The density function of customers is again the constant n. The market share is determined by

$$s = \frac{\left\{ \int_1^C n \, di \right\}}{n}$$

where c is the point where the condition for i_x is no longer valid, ie

$$c < \frac{(2 + a + b_2 + 2G - p_1 + 2s - 2as - b_1 s - b_2 s + p_1 s + p_2 s)}{2}.$$

The market share is then

$$s = \left\{ \frac{(2 \cdot G + a - p_1 + b_2)}{(2a + b_1 + b_2 - p_1 - p_2)} \right\}. \qquad (2.4.i)$$

Consequently, bank 2's market share is

$$\left\{ \frac{(-2 \cdot G + a - p_2 + b_1)}{(2a + b_1 + b_2 - p_1 - p_2)} \right\}.$$

This formula implies effects that are quite intuitive. If the bank charges high fees, its market share declines.

$ds/dp_1 = (2 \cdot G - a + p_2 - b_1)/(2a + b_1 + b_2 - p_1 - p_2)^2$. Whenever formula 2.4.i predicts positive values for bank 2's market share, $ds/dp_1 < 0$, which is reasonable. The impact of b_1 on bank 1's market share is always negative.

If bank 1 is unpopular ($G<0$), its market share remains small. On the other hand, high prices charged by its rival increase its market share.

2.4.2.2 The main case: Bertrand competition outcome with internal point solutions

As mentioned above, the highest possible price equals the reservation price, and the lowest possible price is zero. In this section, it is analysed how a bank sets its prices when neither of these two pricing constraints is binding.

Why do banks charge the same price for both types of payments?

The optimal price for interbank payments is determined according to the first order condition

$$\frac{d\Pi_1}{dp_1} = 0.$$

This condition is satisfied by only one value of p_1, namely

$$p_1 = \frac{2a^2 - \alpha(b_1 + b_2 - p_2) - (b_2 + 2G)(b_1 - b_2 + p_2) - a(2\alpha + b_1 - 3b_2 - 4G + p_2)}{3a - \alpha + b_2 - 2G - 2p_2} \qquad (2.4.ii)$$

When p_1 is given this value,

$$\frac{d^2\Pi_1}{dp_1^2} = -\frac{n(-3a + \alpha - b_2 + 2G + 2p_2)^4}{8(2a + b_2 - p_2)^3(a + b_1 - 2G - p_2)^2}.$$

Whenever $p_2 < 2a+b_2$ holds (as is implied by result 2.4.a and the reaction function 2.4.iii), $d^2\Pi_1 / dp_1^2 < 0$ also holds, which in turn implies that the extreme value is a maximum.

Result 2.4.a
If both prices (p_z, b_z) are between zero and the reservation price, they are equal.

Proof

When p_1 is optimised according to formula 2.4.ii, then

$$\Pi_1 = \frac{n(a + \alpha + b_2 + 2G)^2}{4(2a + b_2 - p_2)}.$$

This implies $d\Pi_1/db_1 \equiv 0$

$\Rightarrow$ The value of b_1 is of no relevance to profits, provided the bank optimizes its p_1 according to formula 2.4.ii.

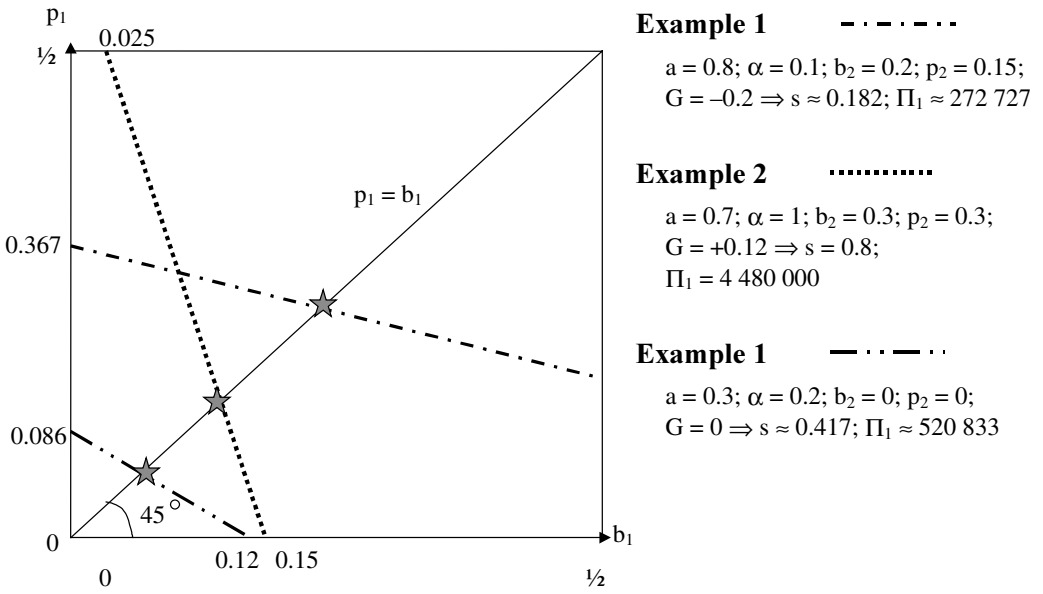
Therefore, in order to avoid the minor administrative cost ε, the bank optimizes the fee for intrabank payments by choosing a combination of prices that satisfies $b_1 = p_1$, whenever such a combination is possible and neither the reservation price constraint nor the non-negativity constraint is binding.

QED

This result does *not* imply that when the price of an interbank payment (p_1) is optimized, any price for an intrabank payment (b_1) would be as good as any other. The price of an intrabank payment (b_1) is one of the factors that affect the optimal price of an interbank payment, and the two prices cannot be chosen independently. There are numerous *different combinations of b_1 and p_1* that would generate the same maximum profit. This maximum profit can be reached via different optimal combinations of p_1 and b_1, but it cannot be exceeded.

Each combination of parameters that is exogenous to the bank's pricing (such as rival's prices, α and a) gives a precisely defined set of combinations of the two prices (p_1 & b_1) that produce the same maximum profit. If one differentiates the expression 2.4.ii with respect to the price b_1, one observes that whenever both banks have a positive market share and they both charge positive prices, the optimal interbank payment fee is linearly dependent on the intrabank payment fee. The price p_1 is negatively related to the price of an intrabank payment. These combinations of the two prices can be presented graphically. A few examples are presented in figure 5. In the following, a set of profit maximizing combinations of p_1 and b_1 will be called a 'pricing curve'.

Figure 5. **Examples of bank 1's pricing curves**



Unless the bank is bound by the reservation price constraint, there is one point on each pricing curve where the two prices are equal. The bank chooses this combination of prices in order to avoid the minor cost ε. In figure 5, these combinations are marked by stars.

The result that a large number of different combinations of the two prices lead to the same maximum profit may sound somewhat counter-intuitive. Nevertheless, the basic idea is farily simple. It is possible to interpret the pricing decision as a two-stage process.

1) First, the bank decides which average price level to offer, whether to be an expensive service provider or to sell services at low prices. This decision is extremely important, and the optimal choice is affected by many different factors, including the expected pricing policy of the rival.

2) Then the bank makes a less important decision in choosing which combination of the two prices to charge in order to implement the chosen average price level. When customers choose the bank, they still do not know with whom they will exchange payments, and their decisions are affected by the expected value of the fee charged by the bank, not by a particular fee. When the bank increases the price of interbank payments and lowers the price of intrabank payments, the average price it offers to its customers may not change at all, at least not in terms of expected value. If the market shares are exactly equal, the changes in the two prices must be equal, because an intrabank payment is as likely as an interbank payment. If, instead, the bank has a small market share, the change in the intrabank price must be much greater to offset a given change in the interbank payment fee, because the latter will be paid by most customers. In figure 5, one can see in the example 2 that a bank with a dominant market share has a nearly vertical pricing curve. The change in the interbank payment fee (p_1) has to be substantial to compensate for a much smaller change in the intrabank fee (b_1), because few customers will pay the interbank payment fee.

In a similar way, the total sum of payment service fees received by the bank does not change, and the bank's profit remains unaffected. In light of this, it is not surprising that the bank's market share is invariant to the point on a chosen pricing curve, provided the bank has chosen a combination of p_1 and b_1 that satisfies the profit maximisation condition. (Proof not shown here.)

Moreover, the profit of rival bank 2 is invariant to the point on the pricing curve chosen by bank 1. As long as bank 1 does not change the average price it charges, there is no change from the point of view of the market structure. (Proof not shown here.)

It is possible to demonstrate that a similar 'irrelevance effect' would obtain for net interest income if it were endogenized, assuming zero price elasticity of demand with respect to financial intermediation.[7]

With formula 2.4.ii, we can calculate that, when the two prices are equal, the optimal pricing is

$$p_1 = b_1 = \frac{(a - \alpha + 2G + b_2)}{2} \qquad (2.4.iii)$$

There are several factors that may cause the bank to charge high prices. These include a well functioning interbank system, low net interest revenue per customer, popularity and the charging of high prices by the rival, all of which are consistent with intuition. Interestingly, the optimal pricing policy depends on the intrabank payment fees charged by the rival but not on its fees for interbank payments.[8]

---

[7] Let δ_z be the net interest income, a transfer of wealth from the customer to bank z ($\alpha = \delta$ – the fixed cost per customer). The value of δ is freely chosen by bank z, but its value is limited by the competitive pressure. Let $p_z = b_z$.

Customer n chooses bank 1 if

$$G + (2 - i_n) + (1-s)\cdot(a - p_1) + s\cdot(1 - p_1) - \delta_1 > -G + (i_n - 1) + s\cdot(a - p_2) + (1-s)(1-p_2) - \delta_2$$

which gives $s = (a - \delta_1 + \delta_2 + 2G - p_1 + p_2)/(2a)$.

There are numerous different combinations of p_z and δ_z leading to the same market share. Let $T_1 = p_1 + \delta_1$. Then the bank's total income equals $n\cdot(a + \delta_2 + 2G + p_2 - T_1)/(2a)\cdot T_1$. The profit does not depend on the combination of p_1 and δ_1 used to implement any chosen value of T_1.

[8] The fact that the rival reacts to only interbank payments has the following implication. If either of the two banks could pre-commit itself to a certain pricing policy, it would no longer be optimal for it to charge the same price for both types of payments. By committing itself to charging different prices for the two payments, bank 1 could affect the pricing decisions of bank 2, which might be optimal for bank 1. The case of pre-commitment in pricing is not analyzed in this paper.

When both banks have committed themselves to charging the same price (bank z's price denoted p_z) for both interbank and intrabank payments, the price will be determined according to the function 2.4.iii. Then, formula 2.4.ii holds as an identity.

When both banks optimize their service fees according to the expression 2.4.iii, prices are strategic complements, as they normally are in Bertrand competition. The outcome of the Bertrand competition is characterized by the following prices, market shares and profits.

Bank 1	Bank 2	
$p_1 = a - \alpha + 2G/3$	$p_2 = a - \alpha - 2G/3$	(2.4.iv)
$s = ½ + G/(3a)$	market share $(1–s) = ½ – G/(3a)$	
$\Pi_1 = n(3a + 2G)^2/(18a)$	$\Pi_2 = n(3a – 2G)^2/(18a)$	

With several different parameter values, the prices predicted by these formulas would lie between 0 and the reservation price level, $(a + \omega)/2$.

An important difference between this result and the previous version of this model is the following. If there are no pricing decisions, as was the case in section 2.3, a small bank cannot try to enter the market by using the price weapon. Because the popular bank will charge relatively high prices whenever it is allowed to do so, a less popular rival can enter the market by undercutting prices. Now, there are three times as many eventual values of G that allow both banks to enter the market. The 'breakeven point' of a small bank is $G = \pm 3a/2$, whereas in the absence of price competition in section 2.3, the break even point was $G = \pm a/2$.

Again, cases where the market share of either of the two banks is close to zero are not realistic. Due to minimum capitalization requirements, such very small banks would not be allowed to enter the market.

2.4.3 The Bertrand competition outcome with binding constraints

2.4.3.1 The outcome with one binding nonnegativity constraint

In the real world, payment services are often cross-subsidized, ie used to attract customers rather than as a significant source of revenue as

such. Thus it is reasonable to study cases where no fees are charged, even though banks would be allowed to charge them. This section focuses on cases where one of the two banks does not charge a fee for making a payment even though its rival does.

Formula 2.4.ii for the optimal interbank payment fee predicts negative values for p_1 with many different parameter values, especially when the net interest income per customer (α) is high. In such cases, it is of paramount importance for a bank to offer an attractive package of payment services in order to gain a maximal market share, because a large market share as such implies high revenues. In practice, payment services can be offered free of charge. Even relatively low values of α often lead to free payment services.

Whenever it is profitable not to charge a fee for an interbank payment, it would be reasonable to offer intrabank payments free of charge as well. There is no reason why a bank would implement a positive average price with a combination consisting of a positive price and a zero price.

The reaction functions of the banks are

$$p_2 = b_2 = \max\left\{\frac{(a - \alpha - 2 \cdot G + p_1)}{2}, 0\right\};$$

$$p_1 = b_1 = \max\left\{\frac{(a - \alpha + 2G + p_2)}{2}, 0\right\}.$$

These pricing rules imply that cases where $p_1 = 0$ and $p_2 > 0$ will not be observed unless bank 2 is more popular than bank 1 ($G < 0$), and the fixed net income per customer (α) is not excessively high. In such cases, the unpopular bank 1 would not charge fees, because charging them would imply a disastrous loss of market share. Bank 2 has insufficient incentive to offer payment services free of charge because it can charge reasonable prices and still maintain a sufficient market share.

To be more precise, the following four conditions must be satisfied to make a reasonable duopoly case where bank 1 does not charge a fee while bank 2 charges a positive prices.

1) $p_2 = (a - \alpha - 2G)/2 > 0$ if $G < (a - \alpha)/2$;
2) $p_1 = 0$ if $G \leq (a - \alpha + p_2)/2 \Leftrightarrow G < (-3a + 3\alpha)/2$.
3) These pricing policies imply that bank 1 has the market share $s = (3a - \alpha + 2G)/(4a)$. $\Rightarrow$ The situation is a duopoly if $(-3a + \alpha)/2 < G$. (If $G < (-3a + \alpha)/2$, bank 2 is a monopoly.)

4) Unless $\alpha > 0$, bank 1, which does not charge any fees, cannot make a positive profit and would not enter the market.

These conditions cannot be satisfied simultaneously unless bank 1 is less popular than its competitor ($G < 0$).

If the parameter G is to satisfy these conditions, the lower bound to G is determined by the third condition; unless G has a certain minimum value, there is a monopoly. If G is given higher and higher values, its upper bound is determined by either of the following two conditions:

– The first condition: $G < (a - \alpha)/2$. With a sufficiently high value of G, both banks offer free services.
– The second condition; $G < -3(a - \alpha)/2$. With a sufficiently high value of G both banks charge fees.

If all the four conditions are satisfied, bank 1 will offer free services and bank 2 will charge fees. The profit for bank 1 will be

$$\Pi_1 = \frac{\alpha(3a - \alpha + 2G)n}{4a}. \qquad (2.4.v)$$

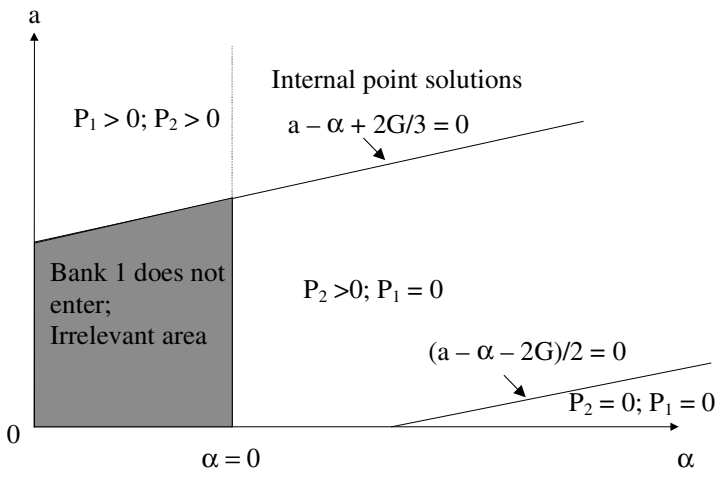
Unsurprisingly, if the fixed net income per customer (α) approaches zero, bank 1 cannot make a profit.

In this case, bank 2 charges positive prices. When it has optimized its prices, its market share equals $1 - s = (a + \alpha - 2G)/(4a)$, and bank 2's profit is

$$\Pi_2 = \frac{n(a + \alpha - 2G)^2}{8a}. \qquad (2.4.vi)$$

Because the bank 2 charges fees for its payment services, its profit does not approach zero as the fixed net income per customer (α) approaches zero.

Figure 6. **Banks' pricing with different values of a and α with G < 0**



Even when reservation prices are assumed not to bind, there are several different combinations of pricing policies. Figure 6 hopefully sheds some light on how banks' pricing decisions depend on two variables, a and α. As can be seen in the figure, a high net income per customer (α) makes banks unwilling to charge for payments, whereas a well functioning interbank system has the opposite effect. If the exogenous net income per customer is very high, neither of the banks is willing to sacrifice any market share in order to earn fee revenue by pricing payment services above zero, thus, $p_1 = p_2 = 0$. If, instead, the quality of interbank payments is high, market shares do not react strongly to prices, and both banks prefer to charge. If both variables have moderate values, the popular bank can charge positive prices and still maintain its position in the market, whereas its rival must price at zero.

If the exogenous net income per customer (α) is negative, a bank must be able to maintain a positive market share even if it charges positive prices; otherwise it would not be able to cover the cost of having customer relationships, and it would not prefer to enter the market.

2.4.3.2 Charging the reservation price

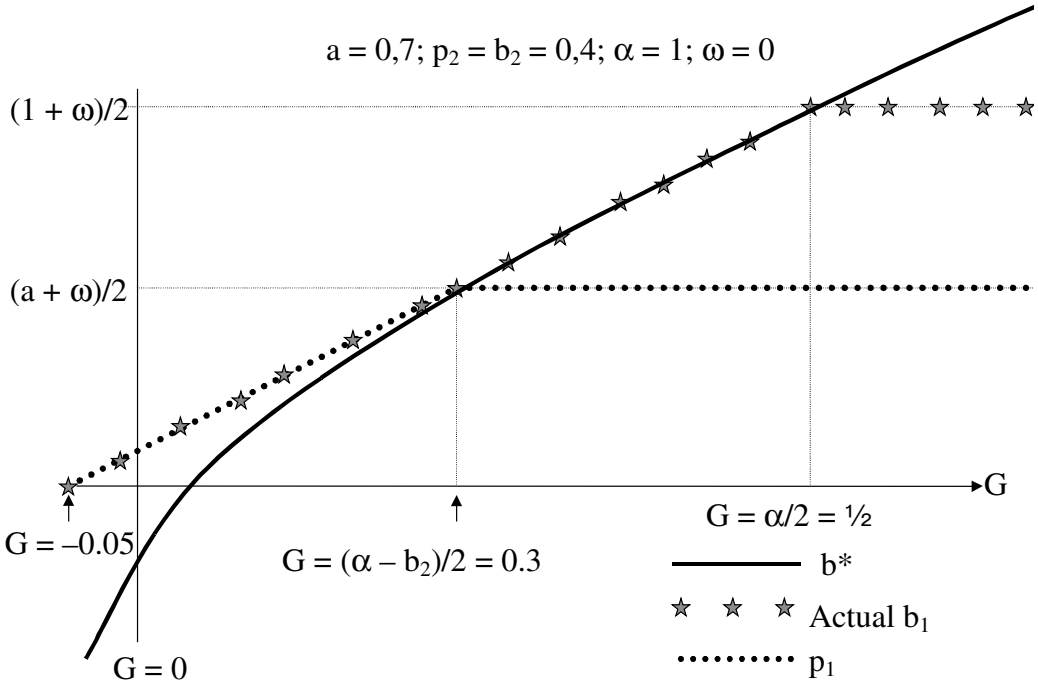
In principle, there are numerous different cases where the reservation price constraint is binding. A bank can be bound by either the reservation price for interbank payments or by both reservation price constraints.

Figure 7 describes how the pricing constraints affect the banks's pricing decisions. If the bank is unpopular, it cannot afford charging anything, because it would lose most of its customers. If it becomes more popular (G increases), it can charge higher and higher prices, as implied by formula 2.4.iii.

At a certain point {when $G = (\alpha - b_2)/2$}, the reservation price constraint for interbank payments becomes binding. It becomes optimal to set the price of interbank payments (p_1) at the reservation level $(a+\omega)/2$, and the price of intrabank payments will be determined according to the function b*, which is the optimal value of b_1 if p_1 is exogenously set at the reservation price.

With a certain value of G, even the value of b* exceeds the reservation price, b_1, and the bank must set both prices at the reservation level.

Figure 7. **Different pricing rules with different values of G: – an example**



It is also possible to construct examples where both banks' pricing decisions have at least one constraint.

In the following, the focus is on cases where both banks set both prices at the reservation level. The most important reason for this is that cases where only one, two or three of the four prices of the Bertrand game are constrained are not, in practice, solvable mathematically.

When all four prices are set at the maximum value [$p_1 = p_2 = (a + \omega)/2$ and $b_1 = b_2 = (1 + \omega)/2$], bank 1's market share is $s = (1 + a + 4G)/(2 + 2a)$. If $G = (-1 - a)/4$, then $s = 0$. If $G = 0$, then $s = ½$, and if $G = (1 + a)/4$, then $s = 1$.

The respective profits are

$$\Pi_1 = \frac{n(1+a+4G)\left[1+a^2+4\alpha+4G+2\omega+2a(1+2\alpha-2G+\omega)\right]}{8(1+a)^2} \quad (2.4.vii)$$

and

$$\Pi_2 = \frac{n(1+a-4G)\left[1+a^2+4\alpha-4G+2\omega+2a(1+2\alpha+2G+\omega)\right]}{8(1+a)^2}. \quad (2.4.viii)$$

2.4.3.3 Sabotage pricing

One potential outcome of the model is that a bank with a large and dominant market share would be able to make its competitor an unattractive choice for customers. If the dominant bank charges excessively high prices for interbank payments, its customers will use cash in payments to customers of the competing bank. This would affect the utility of payer and payee equally. In this hypothetical case, a typical customer of the smaller bank would receive payments for the most part from customers of the large bank, whereas a typical customer of the large bank would not be severely affected. Only a minor portion of the latter's payments would go to customers of the small bank. Thus such a pricing policy would affect much more the average quality of payment services of a customer of the small bank than the quality of service perceived by a customer of the large bank.

Cases where sabotage pricing is profitable can be found if the disutility of cash payments (ω) is high and the quality of the interbank payment system (a) is good. Even though these cases are in principle possible, they will not be analysed in detail.

2.5 Investment when banks Bertrand-compete

2.5.1 The main case: Neither of the two banks bound by constraints

2.5.1.1 Actual development efforts

When banks charge a fee for a payment service, the value of a affects banks' profits in two different ways:

1) As in section 2.3, it affects market shares. As to this effect, this version does not contain many new properties as compared to the version where prices were assumed to equal zero. The smaller bank invests in the system in order to increase its market share, whereas its big rival has no such incentive.

2) It affects the price banks can charge for a payment service. This effect could not have existed in the model of section 2.3, but it is implied by the reaction function (2.4.iii).

Because the second effect could not obtain in the absence of service fees, it is worth closer scrutiny.

Even though the demand for payment services is fixed and exogeneous, the quality of payment systems affects the equilibrium price (see 2.4.iii). This may sound counterintuitive, but the result has a reasonable explanation. If interbank payments are slow and unreliable, it is highly important for customers to use the same bank as the majority. And the larger the majority, the less attractive it is to belong to the minority. If a bank manages to attract customer A from its rival by offering payment services at a low price, customer B might follow suit, because he may have to exchange payments with A. Thus market shares overreact to prices.

If instead interbank payments function well, the use of low prices as a competitive weapon has no cumulative effects, because it is no longer essential for customers to use the same bank as the majority. Gaining an additional customer has no cumulative effects. This result might be valid even more generally; *in the presence of direct network externalities, improving the compatibility of networks might relax price competition*.

This result has analogies with the model of Katz and Shapiro (1986). This is a discrete time model describing two competing firms selling at consecutive stages products characterized by direct network

externalities. Each firm must make a one-off choice between supplying compatible or incompatible products. If products are incompatible, network externalities are limited to customers of the same company. All consumers have identical preferences, implying that the winning firm finally gets all the customers if technologies are incompatible. A central concept in the model is the so-called 'installed base advantage', ie a product that was widely used in the past yields higher utility now. At the early stage of competition, firms supplying incompatible products compete fiercely with prices in order to provide customers with superior network externalities. In the beginning, incompatible products can even be sold at a loss at subsidized prices. Gaining the position as dominant firm is essential to future (monopoly) profits. Product compatibility relaxes competition in the short term but intensifies it in the long term, because compatibility enables both firms to be permanently present in the market.

This finding may seem to be inconsistent with the results of Matutes and Padilla (1994). They demonstrated that in a spatial oligopoly, compatibility of ATM networks intensifies price competition between banks. If ATM networks are compatible, customers can make transactions through all ATMs, and the monopoly power provided by location and geographic distance partly disappears. The difference between the result of Matutes & Padilla and that presented above is due to the entirely different role of compatibility in the system. In the model of Matutes & Padilla, improving the compatibility erodes market power, because compatibility reduces the relative importance of geographic distance. In this model, compatibility has no impact on access to the services of the two banks. This makes it less important for a customer to use the same bank as the majority. Hence compatibility actually *increases* the relative significance of the location parameter (i).

The profit function of a bank and its dependence on the quality of interbank payments can be characterized as follows. If both banks charge positive prices that are not bound by the reservation price constraint, then, according to (2.4.iv), $\Pi_z = n(3a + 2G)^2/(18a)$, z = 1,2.

$$\Rightarrow d\Pi_z / da = n\left(9a^2 - 4G^2\right)/\left(18a^2\right).$$

Result 2.5.a
If a bank has a market share between 0 % and 100 % (which will be the case iff –3a/2 < G < +3a/2), it will invest in system development. If its market share is 0 % or 100 %, it will not invest.

Proof

If the market share of bank 1 is 100 %, then $s = ½ + G/(3a) = 1$, so that $G = 3a/2$.

Therefore $d\Pi_1/da = n(9 - 4G^2/a^2)/18 = n(9 - 4\cdot(3a/2)^2/a^2)/18 = 0 \Rightarrow$ The bank has no incentive to invest in system development.

If the market share is 0 %, then $s = ½ - G/(3a) = 0$ so that $G = -3a/2$. Therefore $d\Pi_1/da = n[9 - 4\cdot(-3a/2)^2/a^2]/18 = 0 \Rightarrow$ The bank has no incentive to invest in the system.

If $0 < s < 1$, then $|G| < 3a/2$, which implies $d\Pi_1/da = (n/18)\cdot(9 - 4G^2/a^2) > 0 \Leftrightarrow 9a^2 > 4G^2$, which is true whenever $|G| < 3a/2$.

Because the marginal impact of investment in payment system quality is infinite if nothing has been previously invested in the system, it cannot be optimal for the bank to invest nothing in the system when $-3a/2 < G < +3a/2$.

QED

Unlike in the case of free payment services, a bank with a zero market share does not invest in system development. An unpopular bank that has difficulties in finding any customers uses low prices to attract customers rather than socially undesirable excessive investment in the payment system.

Investments by a small bank would help the bank both to gain a positive market share and to charge higher prices, but, paradoxically, the bank would not be interested in entering the market. The pricing formula (2.4.iii) does not predict that a bank with an almost zero market share would charge positive prices unless the fixed net income per customer (α) is negative and G is close to $\pm 3a/2$. Hence the difference between this result (2.5.a) and result 2.3.b is not simply due to the fact that payment services are now assumed to be costly. Instead, this analysis applies to cases that cannot be meaningfully analysed if one assumes that no fees are charged. As concluded above, there is now a much wider range of different values of G that lead to a duopoly situation where both banks have a nonzero market share.

When a bank has a market share of about 50 %, improvements in the payment system do not affect its customer base negatively. Instead, it can collect fees for a substantial amount of payment

services. Thus it is not surprising that a medium-sized bank has the strongest incentive of all to invest in the system.

Result 2.5.b
Private expenditure on payment system development as a function of G reaches its maximum when G = 0.

Proof:

The optimization condition for bank 1 is $(d\Pi_1/da)\cdot(da/d\Lambda_1) - 1 = 0$. Implicit differentiation gives

$$\frac{d\Lambda_1}{dG} = -\frac{\left(\frac{d^2\Pi_1}{dGda}\right)\cdot\left(\frac{da}{d\Lambda_1}\right)}{\frac{d^2\Pi_1}{d\Lambda_1^2}}.$$

Optimization implies that $d^2\Pi_1/d\Lambda_1^2 < 0$ and by assumption $da/d\Lambda_1 > 0$.

Hence, $d\Lambda_1/dG = 0$ if $d^2\Pi_1/dGda = 0$.

$d^2\Pi_1/dGda = -n(4/9)\cdot G/a^2$, which cannot be 0 unless $G = 0 \Rightarrow$ There is only one extreme value in investment as a function of G.

The second order condition is:

$$\frac{d^2\Lambda_1}{dG^2} = -\frac{\left(\frac{d^3\Pi_1}{dG^2da}\right)\left(\frac{da}{d\Lambda_1}\right)\left(\frac{d^2\Pi_1}{d\Lambda_1^2}\right) - \left(\frac{d^3\Pi_1}{dGd\Lambda_1^2}\right)\left(\frac{d^2\Pi_1}{dGda}\right)\left(\frac{da}{d\Lambda_1}\right)}{\left[\frac{d^2\Pi_1}{d\Lambda_1^2}\right]^2}$$

where $d^3\Pi_1/dG^2da = -(4/9)/a^2 < 0$. $da/d\Lambda_1 > 0$ and $d^2\Pi_1/d\Lambda_1^2 < 0$. When $G = 0$, $d^2\Pi_1/dGda = 0$.

It follows that $d^2\Lambda_1/dG^2 < 0$ and the extreme value is a maximum.

QED

The most surprising result may be the following, which is called the *symmetric incentives property*.

Result 2.5.c
If both banks charge a price that lies between the reservation price and zero, they will always spend an equal amount in system development ($\Lambda_2 = \Lambda_1$).

Proof:

Bank's profit is $\Pi_1 = n\cdot(3a + 2\cdot G)^2/(18a)$.

Analogically, Bank 2's profit is $\Pi_2 = n\cdot(3a - 2\cdot G)^2/(18a)$.

Differentiation with respect to a yields $d\Pi_1/da = (n/18)\cdot(9 - 4G^2)/(a^2)$ and $d\Pi_2/da = (n/18)\cdot(9 - 4G^2)/(a^2)$, which are equal; $d\Pi_1/da = d\Pi_2/da$.

The incentive for investing in the system is therefore always the same for both banks, and both banks invest the same amount in system development.

QED

This symmetric incentives property can be understood intuitively as follows: As in the model without payment service fees, the smaller bank can gain a larger market share by investing in the system. The big bank, by contrast, benefits in absolute terms much more than the small one from the impact of payment system development on the price of the service. The impact of development efforts on market shares is unfavourable, but the price effect more than offsets this negative effect. Overall, the big bank benefits as much as the small bank.

The symmetric incentives property has several interesting implications. For instance, it implies that the banks always react similarly to different exogenous factors that might affect the optimal amount of investment.

Due to the property of symmetric incentives, the resulting quality of the payment system (a) is reached in a cost efficient way. The assumptions concerning the a-function imply that any given value of *a* is achieved in the most cost-efficient way if both private banks invest the same amount in system development. And when both banks have equally strong incentives to invest, their investments are equal.

In many duopoly models, it is interesting to know whether decision variables are strategic substitutes or complements. In this model, there is no generally valid answer to the question. In many cases, investments would be strategic substitutes, but this result is not

always true. If investments by the two private agents have a negative interaction effect on the quality of the system ($d^2a/d\Lambda_2 d\Lambda_1 < 0$), the investments are strategic substitutes with many different parameter values. If they have a positive interaction effect, they are always strategic complements. However, whether investments are strategic substitutes or complements is of minor importance for optimal central bank investment.

2.5.1.2 Actual vs socially optimal investments

First, we shall review how payment system development affects the utility of customers. The welfare of a bank 1 customer equals

$$W_x = [G + (2 - i_x) + (1 - s) \cdot (a - p_1) + s \cdot (1 - p_1)]$$

and the impact of payment system development on the welfare of bank 1 customers equals = $1 - s - dp_1/da + ds/da \cdot (1 - a)$.

The welfare of a bank 2 customer equals $W_x = -G + (i_x - 1) + s \cdot (a - p_2) + (1 - s) \cdot (1 - p_2)$ and the impact of a on the welfare of a bank 2 customer is $s - dp_2/da - (ds/da) \cdot (1 - a)$.

The impact of a change in *a* on consumer welfare in the whole economy can be calculated as in the case of free payment services (see 2.3.iii). The total impact equals

$$n \cdot s \cdot \{1 - s - dp_1/da + ds/da \cdot (1 - a)\} +$$
$$n \cdot (1 - s) \cdot \{s - dp_2/da - (ds/da) \cdot (1 - a)\} = -n[½ + 2G^2/(9 \cdot a^3)].$$

Interestingly, the impact of payment system development on consumer net welfare is negative. Although this may seem counter-intuitive, there is a natural explanation for it. The effect is basically due to the fact that price competition between banks is relaxed by improvements in the payment system, which is certainly undesirable from the consumer's point of view. As a whole, the price of a payment service increases proportionately with the quality of interbank payments. When the quality of interbank payments improves, intrabank payments do not improve, but the customer has to pay more even for them. Therefore, payment system development implies a transfer of wealth from customers to banks.

When the market is highly concentrated, an improvement in the quality of the payment system is of little use to customers, because few payments are processed through the interbank system. A well

functioning system actually discourages the smaller bank from engaging in aggressive price competition, even though competition would benefit consumers. Thus, in a concentrated market, payment system improvement is especially undesirable for customers.

Needless to say, the impact of payment system development on gross consumer utility prior to payment of service fees, is positive for many different parameter values.

Result 2.5.d
The total impact of payment system improvement on social welfare is positive ($d\psi/da > 0$) iff $|G| < 3a/(2\cdot\sqrt{[2 + 1/a]})$.

Proof:

It was demonstrated in the result 2.5,c that the impact of the quality of interbank payments (a) on profits equals $\{(n/18)\cdot(9 - 4G^2/a^2)\}$.

Moreover, it was demonstrated in the beginning of this section that the impact of the quality of interbank payments on the consumer utility equals $-n[½ + 2G^2/(9\cdot a^3)]$.

Thus $d\psi/da = -n[½ + 2G^2/(9\cdot a^3)] + 2\cdot\{(n/18)\cdot(9 - 4G^2/a^2)\} = n(9a^3 - 4G^2 - 8aG^2)/(18a^3)$, which is positive iff $|G| < 3a/(2\cdot\sqrt{[2 + 1/a]})$.

QED

When a bank invests in system development, the resulting improvement has several consequences. First, the investment affects the equilibrium price of payment services, thereby causing transfers of wealth from customers to the banking industry. These transfers of wealth do not cause any allocative distortions, and they are harmless from the point of view of social welfare. Secondly, customers receive improved interbank payment services, which is a positive effect. And finally, the investment affects market shares, which in most cases is undesirable, because it hampers the functioning of the payment system by increasing the market share of the smaller bank.

Result 2.5.e
Unless each bank has a 50 % market share ($G = 0$) or either of the banks has a 0 % market share ($|G| = \pm 3a/2$), the banks invest more than the social optimum in payment system development.

Proof:

The total externality of payment system development by bank 1 equals the sum of the impact on the rival's profits: $(n[½ – 2G^2/9a^2])$ and the impact on consumer welfare $(–n\cdot[½ + 2G^2/(9\cdot a^3)])$.
The total externality equals $= –2n(1 + a)G^2/(9\cdot a^3)$
Unless G = 0, this is negative, and private benefits exceed public benefits.

In addition, if $G = \pm 3a/2$, the bank invests nothing in the payment system, and the externalities of a hypothetical investment would be negative because $|G| > 3a/2\cdot\sqrt{[2 + 1/a]}$. There would be a disparity between the private and social benefits of eventual investments. Nevertheless, there would be no disparity between the actual and socially optimal level of investment. When $G = \pm 3a/2$, the market shares are (100 %, 0 %) and, according to 2.5.a, there is no private investment. As has been seen (2.5.d), zero investment in the payment system would be desirable. In this case, there is no private investment, which is a socially optimal outcome.

QED

Hence, as a rule, banks overinvest in system development. As demonstrated in the previous result, a bank causes transfers of wealth from customers to both itself and its rival by investing in the system. Moreover, by investing in the system, a small bank can also increase its market share at the cost of the rival, which is also socially useless and even harmful.

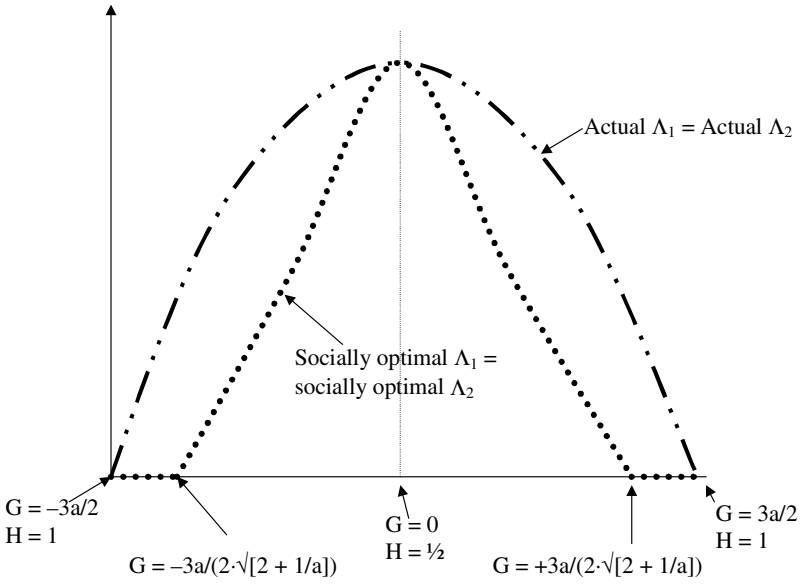
When the two banks are equally popular, investing in the system does not affect market shares. However, the investment still has three effects on other sectors of the economy:

1) A transfer of wealth from customers to investing bank
2) A gross increase in consumer utility yielded by improved payment services
3) A transfer of wealth from customers to the rival bank, caused by the increase in service fees. This is a mere transfer of wealth, which as such does not affect social utility.

When market shares are equal, effects 1 and 2 offset each other and, in the aggregate, there is no net externality for the rest of the economy.

Private and public benefits are equal, and private investment is at the socially optimal level. Interestingly, it is not possible to construct examples where the private sector would invest less than the socially optimal amount.

Figure 8. **Optimal and actual private investments**



If the two banks were to cooperate in payment services, the outcome might be worse. This is an implication of the symmetric incentives principle. Both banks would always prefer a higher level of rival investment. An additional investment by rival 2 would increase the profit of bank 1, but the costs would be borne by rival 2. If the two banks cooperated, they would agree on increasing their investment expenditure further. Hence there would be more over-investment.

One of the main conclusions seems to be that banks rather systematically over-invest in the interbank system, compared to the social optimum. Needless to say, the comparison has been made between the actual outcome and the level of investment that would be optimal given banks' profit maximizing behaviour in terms of price competition. This does not automatically imply that the actual level of investment would be systematically higher than in the first best solution with almost any given value of G. What if the two banks tried to set prices so as maximize social welfare instead of profits?

Because there is a reservation price instead of an ordinary price elasticity of demand, the situation entails certain particularities. If both banks cut or increase their prices in a similar way, this would simply

lead to a transfer of wealth between customers and banks. Thus, providing payment services free of charge might not be optimal even though the marginal cost of producing payment services is zero. Pricing decisions affect market shares. The small bank would charge prices whereas its large competitor would not, because it would be socially optimal to increase the market share of the large bank in order to reduce the number of interbank payments. Market shares would react to different values of G even more strongly than in the case of free payment services. Thus, with any given non-zero value of G, the socially optimal quality of interbank payments would be even weaker than what it is if banks were to price their payment services so as to maximize profits, because market shares would differ more from the 50 %–50 % situation than in the actual outcome. Hence, if one compares the actual outcome with the first-best situation, the over-investment is probably even more extreme than if banks' profit maximization in the Bertrand competition is taken for granted.

2.5.1.3 Optimal central bank policies

Again, central bank investment in the payment system has two kinds of effects. First, it directly affects the quality of payment services. Secondly, it has an indirect effect through the reactions of the private sector. The symmetric incentives property significantly simplifies the analysis of the indirect effects.

Result 2.5.f
Central bank investment affects both banks' investment in a similar way: $\partial\Lambda_1/\partial\Lambda_c = \partial\Lambda_2/\partial\Lambda_c$**.**

> This result is a direct corollary of the symmetric incentives property.

Because now both banks invest and because they react to each other's investments, the total impact of central bank investment on private investment is more complicated than in section 2.3. If investments by the two private banks are strategic complements, the indirect effects *strengthen* the impact of central bank investment on private investment. To take an example, an increase in central bank investment might encourage bank 1 to increase its investment. Due to symmetric incentives, bank 2 would also increase its own investment. And because of strategic complementarity, bank 1 would react to its

rival's increased investment by increasing its own investment further. If instead investments are strategic substitutes, the situation is different, and the indirect effect through the reaction of bank 2 would weaken the total impact of central bank investment on bank 1's behaviour.

Nevertheless, qualitative conclusions concerning the impact of central bank investment on private investment are not reversed by indirect effects. If the increase in central bank investment provides a (dis)incentive for a bank to invest in developing the system, rival reactions are not strong enough to reverse this effect.

Result 2.5.g

Let $\partial\Lambda_z/\partial\Lambda_c$ denote the direct impact of central bank investment on private investment, where rival's investment is exogenous. Let $d\Lambda_z/d\Lambda_c$ denote the actual total impact, taking into account all direct and indirect effects on investment by the two banks. If the direct impact of central bank investment on private investment, Λ_z, is negative, then in no stable subgame perfect equilibrium can indirect effects through the rival's reactions reverse the impact. ($\partial\Lambda_z/\partial\Lambda_c < 0 \Rightarrow d\Lambda_z/d\Lambda_c < 0$). And iff $\partial\Lambda_z/\partial\Lambda_c > 0$, then $d\Lambda_z/d\Lambda_c > 0$.

Proof:

See appendix 4.

This result simplifies the analysis concerning optimal central bank investment. Because the private sector typically invests more than the socially optimal amount (result 2.5.e), it is reasonable for the central bank to try to reduce private investments by adapting its own investment behaviour. Therefore, it is of special interest to know whether central bank investment can reduce private investment and, if so, under which circumstances. It turns out that it is not possible to reduce private investment by increasing public investment unless $d^2a/d\Lambda_z d\Lambda_c < 0$ and $\Lambda_c > 0$. Even in such cases, it is not certain that central bank investment would actually reduce private investment.

The easiest way to handle this problem mathematically is to analyse the possibilities of the central bank to do the opposite, ie to increase private investment, which would actually never be optimal. The reason for this is simple: it is possible to specify sufficient conditions for $d\Lambda_z/d\Lambda_c > 0$.

Result 2.5.h
The total impact of central bank investment on private investment is positive ($d\Lambda_z/d\Lambda_c > 0$) if either

$\partial^2 a/\partial\Lambda_z\partial\Lambda_c \geq 0$ (case 1)

or

$\Lambda_c \approx 0$ and $G \neq 0$ (case 2)

or both.

Proof:

The f.o.c of bank 1 is $\partial\Pi_1/\partial\Lambda_1 - 1 = 0$.

Implicit differentiation gives

$$\frac{\partial\Lambda_1}{\partial\Lambda_c} = -\frac{\left[\frac{\partial^2\Pi_1}{\partial\Lambda_1\partial\Lambda_c}\right]}{\left[\frac{\partial^2\Pi_1}{\partial\Lambda_1^2}\right]}.$$

Calculating the expression for $[\partial^2\Pi_1/\partial\Lambda_1\partial\Lambda_c]$ yields

$$= -\frac{8G^2\cdot\left(\frac{\partial a}{\partial\Lambda_c}\right)\left(\frac{\partial a}{\partial\Lambda_1}\right) + a\left(+9a^2 - 4G^2\right)\cdot\left(\frac{\partial^2 a}{\partial\Lambda_1\partial\Lambda_c}\right)}{\left[\frac{\partial^2\Pi_1}{\partial\Lambda_1^2}\right]\cdot 18a^3}.$$

The second order optimization condition for the bank implies that the denominator is always negative, which implies that the whole term is negative if the numerator is.

Because with any parameter values $8G^2\cdot(\partial a/\partial\Lambda_c)(\partial a/\partial\Lambda_1) > 0$, and because in the duopoly case $+9a^2 - 4G^2 > 0$, the numerator cannot be negative unless $(\partial^2 a/\partial\Lambda_1\partial\Lambda_2 < 0) \Rightarrow$ If $\partial^2 a/\partial\Lambda_z\partial\Lambda_c \geq 0$, public investment always increases private investment (case 1).

If Λ_c approaches zero, then $(\partial a/\partial\Lambda_c)$ approaches $+\infty$, and unless $G = 0$, $8G^2\cdot(\partial a/\partial\Lambda_c)(\partial a/\partial\Lambda_1) = +\infty$. $\Rightarrow$ If both $G \neq 0$ and Λ_c is close to zero, then the numerator must be positive (case 2).

According to the 2.5.g $\partial\Lambda_1/\partial\Lambda_c > 0 \Rightarrow d\Lambda_1/d\Lambda_c > 0$.

QED

Expressing this result verbally in a more intuitive way, if public investment strengthens the effects of private investment on the quality of payment systems ($\partial^2 a/\partial\Lambda_c\partial\Lambda_z > 0$), it is not surprising that central bank investment is always a stimulus to private investment.

At very low levels of central bank investment, the impact of public investment on private investment is always positive. A marginal increase in central bank investment always encourages the private sector to increase its own investment. The private bank may be discouraged from investing if the quality of the existing system is so poor that no major improvements in the market situation can be achieved at a reasonable cost. Thus, even when central bank investment makes private investment technically inefficient, it is not obvious that private investment would become unprofitable for the bank if the central bank increases its investment.

Nevertheless, a certain amount of public investment is justified if the market shares are roughly equal. The investment may have undesirable effects on the behaviour of the private sector, but when there are a lot of interbank payments, the direct benefits of public investment more than offset the undesirable indirect effects.

Result 2.5.i
It is not optimal for the central bank to invest in developing the system if $|G| \geq 3a/(2\cdot\sqrt{[2 + 1/a]})$; if G = 0, it is optimal to invest.

Proof:

See appendix 5.

As a rule, banks overinvest in the system. Therefore, the central bank should try to reduce private investment. Whenever the marginal impact of central bank investment on private investment is negative (positive), it is reasonable for the central bank to increase (decrease) its investment in order to reduce private investment, except when there is no market failure because the market shares are equal (G = 0).

Result 2.5.j
If $G \neq 0$ but it is optimal for the central bank to invest, then the following condition holds. If either $\partial^2 a/\partial\Lambda_1\partial\Lambda_c \geq 0$ or $\Lambda_c \approx 0$ or both, then the central bank should restrict its investment in order to reduce private investment.

Proof:

The optimization condition for the central bank is

$$\frac{d\psi}{d\Lambda_c} = \left(\frac{d\psi}{da}\right)\left(\frac{\partial a}{\partial \Lambda_c}\right) - 1 + \left[\left(\frac{d\psi}{da}\right)\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right) - 1\right]\cdot\left(\frac{\partial \Lambda_1}{\partial \Lambda_c}\right) + \left[\left(\frac{d\psi}{da}\right)\cdot\left(\frac{\partial a}{\partial \Lambda_2}\right) - 1\right]\left(\frac{\partial \Lambda_2}{\partial \Lambda_c}\right) = 0.$$

If $|G| < 3a/(2\cdot\sqrt{[2 + 1/a]})$ but $|G| > 0$, then $(d\psi/da)\cdot(da/d\Lambda_z) - 1 < 0$.

(Result 2.5.e) If either $\partial^2 a/\partial\Lambda_z\partial\Lambda_c \geq 0$ or $\Lambda_c \approx 0$, then $d\Lambda_z/d\Lambda_c > 0$.

(Result 2.5.h) $\Rightarrow$ $[(d\psi/da)\cdot(\partial a/\partial\Lambda_1) - 1]\cdot(\partial\Lambda_1/\partial\Lambda_c) < 0$ and $[(d\psi/da)\cdot(\partial a/\partial\Lambda_2) - 1](\partial\Lambda_2/\partial\Lambda_c) < 0$.

The optimization condition $d\psi/d\Lambda_c = 0$ cannot hold unless $(d\psi/da)(\partial a/\partial\Lambda_c) - 1 > 0$, ie unless the direct impact of central bank investment on welfare is positive, and the value of Λ_c below the level that would be optimal in absence of private sector reactions.

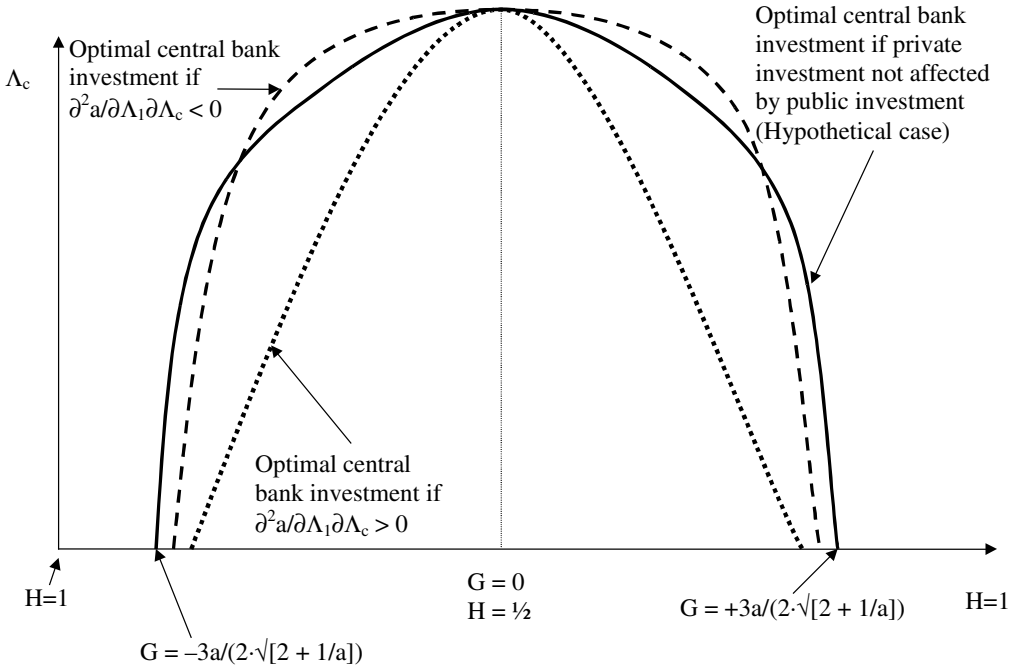
$\Rightarrow$ The central bank should restrict its investment.

QED

However, in some cases the reverse result might hold. If central bank investment is well above zero and public investment makes private sector investment inefficient, the central bank might have good reasons to *increase* its investment to discourage private investment. If the market shares are equal ($G = 0$), the central bank should not try to affect private investment because private investment is already at its desired level.

Optimal central bank policies are illustrated in figure 9.

Figure 9. **Optimal central bank investment when both banks charge for payments**



2.5.2 Investment with binding constraints

2.5.2.1 Investment with a binding nonnegativity constraint

The case where both banks price at zero has already been analysed in section 2.3. It is essentially irrelevant to banks' investment decisions whether prices are zero because of an exogenous constraint that forces banks not to charge or whether the banks deliberately choose to set their prices at zero.

Therefore, in this section, the focus is on banks' incentives to invest in the system when either of them charges for payment services when its rival does not charge. Market outcomes for such cases were analysed in section 2.4.3.1.

Now, it is assumed that

– Bank 1 has a small market share and it offers payment services free of charge in order to increase its clientele.
– Bank 2 charges positive prices.

Both banks have a positive market share, or at least the less popular bank 1 would gain a nonzero market share if the functioning of the interbank payment system (a) were marginally improved.

The impact of payment system development on bank 1's profits equals $d\Pi_1/da = n\alpha(\alpha - 2G)/(4a^2)$, and the impact of payment system development on bank 2's profits equals $d\Pi_2/da = = n\{a^2 - (\alpha - 2G)^2\}/\{8a^2\}$.

It is easy to demonstrate that if bank 2 dominates the whole market, it has no incentive to develop the payment system. This is understandable because developing the system would mainly erode its dominant market position.

Result 2.5.k
If bank 2 has nearly a 100 % market share [which will be the case if $G \approx (-3a + \alpha)/2$], then it cannot be optimal for it to invest in development of the payment system.

Proof:

$d\Pi_2/da = n\{a^2 - (\alpha - 2G)^2\}/\{8a^2\}$; If $G = -(3a - \alpha)/2$, then $d\Pi_2/da = -n < 0$.

The profit decreases if the payment system begins to function better.

QED

As to the case where market shares that are nearly equal (higher values of G), it is difficult to draw any robust conclusions concerning the behaviour of bank 2. Bank 2 may have an incentive to invest in the system with a sufficiently high value of G, but this is not certain. Improvements in the payment system become less and less harmful to bank 2 as its market size decreases ($d^2\Pi_2/dadG = (\alpha - 2G)/2a^2$; $G < 0 \Rightarrow d^2\Pi_2/dadG > 0$), implying that at a certain point improving the interbank system *may* become profitable for bank 2. The improvement would lower its market share, but, it would also improve its possibilities to charge high prices. The effect of such pricing possibilities might eventually more than offset the adverse impact on market share.

Because the marginal impact of a very small investment (Λ_2) on the quality of the payment system (a) is disproportionately strong, bank 2 would invest at least something in the system whenever $d\Pi_2/da$

> 0. However, it is possible to construct examples where $d\Pi_2/da < 0$ for any value of G.

Bank 1, by contrast, has incentives somewhat similar to those in the case where neither of the two banks charges for payments. The bank can increase its market share by making it less burdensome for its own customers to make and receive payments. Moreover, it has an additional incentive. Improving the system encourages bank 2 to charge a higher price (p_2), which also helps bank 1 to increase its market share. Hence it is not surprising that bank 1 always invests in the system.

Result 2.5.l
Bank 1, which does not charge for payments, will invest in the payment system.

Proof:

$d\Pi_1/da = n\alpha(\alpha - 2G)/(4a^2)$.
If $G < 0$, then $d\Pi_1/da > 0$.
When $\Lambda_1 = 0$, then $\partial a/\partial \Lambda_1 = \infty$.
It cannot be optimal for bank 1 not to invest in the system.

QED

Result 2.5.m:
The incentives for bank 1 to invest in payment system development decrease as G increases.

Proof:

$d^2\Pi_1/dadG = -2/(4a_2) < 0$. Thus, the higher the value of G, the less the incentives to invest in the system.

QED

In most cases, bank 1, which charges no fees, invests more in the system than bank 2. Surprisingly, it is also possible to find contrary examples.[9]

Result 2.5.n
If investment by the bank z (z = 1,2) is greater than zero, the impact of central bank investment on private investment, Λ_z, is negative, at least if either $\Lambda_c \approx 0$ or $\partial^2 a/\partial\Lambda_z\partial\Lambda_c \leq 0$.

Proof:

See appendix 6.

The payment services-related consumer surplus of a bank 1 customer equals $U_1 = s + (1 - s)\cdot a$. The impact of a payment system improvement on the consumer surplus of a bank 1 customer equals $dU_1/da = (a_2 + \alpha - 2G)/(4a^2)$. The effect is positive, which is not surprising, because the customer benefits from the increased market share of bank 1, but does not have to pay anything for the improvement. In fact, consumer surplus equal consumer utility.

The payment services-related consumer surplus of a bank 2 customer equals $U_2 = s\cdot(a - p_2) + (1 - s)\cdot(1 - p_2)$. And the impact of a payment system improvement on the consumer surplus of a bank 2 customer equals $dU_2/da = (a^2 - \alpha + 2G)/(4a^2)$, which can be either negative or positive.

The total impact of payment system development on consumer surplus in the whole economy equals

$$\frac{dU}{da} = n \cdot s \cdot \frac{dU_1}{da} + n \cdot (1-s) \cdot \frac{dU_2}{da} = \frac{\left[2a^3 + a(\alpha - 2G) - (\alpha - 2G)^2\right]n}{8a^3}$$

[9] The issue was tested with five thousand simplistic numerical simulations. When a, α and G were given uniformly distributed random values that satisfied the four conditions, bank 2 had a greater incentive to invest in the system than bank 1 in slightly more than 10 % of the cases.

Result 2.5.o
The investment by bank 1 exceeds its socially optimal level at least if bank 1 market share is close to 0, which would be implied by $G = (-3a + \alpha)/2$.

Proof:

The total externality caused by the investment is

$$\left(\frac{dU}{da} + \frac{d\Pi_2}{da}\right)\frac{\partial a}{\partial \Lambda_1} =$$
$$\left[\frac{\left[2a^3 + a(\alpha - 2G) - (\alpha - 2G)^2\right]n}{8a^3} + n\frac{a^2 - (\alpha - 2G)^2}{8a^2}\right] \cdot \left(\frac{\partial a}{\partial \Lambda_1}\right)$$

When $G = (-3a + \alpha)/2$, $dU/da + d\Pi_2/da = -7/8 - 3/(4a) + a^2$.
Whenever $0 \leq a \leq 1$, this is negative.
$\Rightarrow$ bank 1 investment causes a negative externality.

QED

By investing heavily, the small bank can take market share from its rival. This policy both affects customers adversely by increasing the number of interbank payments and reduces the rival's profits. Thus all the effects on other agents in the economy are negative. The central bank can discourage (encourage) bank 1 investment by increasing (decreasing) its own investment, at least if $\partial^2 a/\partial \Lambda_1 \partial \Lambda_c < 0$. {Result 2.5.n}

As already mentioned, the profit maximizing investment by bank 2 is often zero, which is not always socially optimal. In these cases, it is difficult for the central bank to encourage it to invest.

2.5.2.2 Investment when banks price at the reservation level

In this section, it will be analysed how banks invest in the system if they both set their prices at the respective reservation levels. Prices, profits and market shares in such cases were analysed in section 2.4.3.2.

This case has certain analogies with both the case analysed in section 2.3 and the case of internal point solutions. Both banks have

an incentive to invest because a better system enables them to charge higher prices. On the other hand, an improved system would help the smaller bank to gain more market share, which would be beneficial for the small bank but not for the large one.

Mathematically, the situation can be analysed as follows. The impact of improved payment services on bank 1's profits is

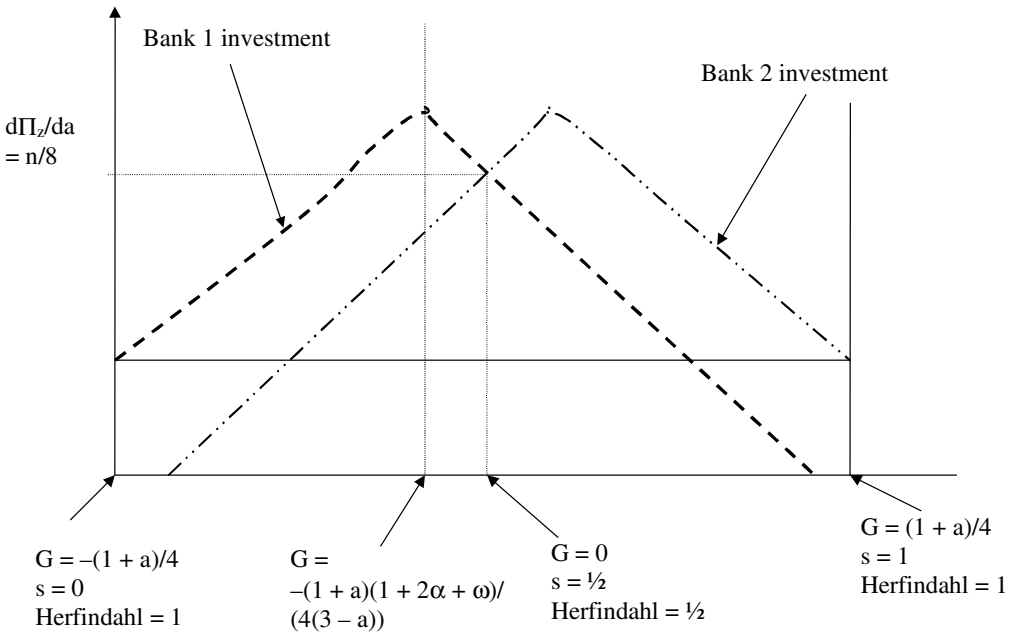
$$\frac{d\Pi_1}{da} = \frac{n\left\{1 + 3a^2 + a^3 - 48G^2 - 8G[1 + 2\alpha + \omega] + a\left[3 + 16G^2 - 8G(1 + 2\alpha + \omega)\right]\right\}}{8(1 + a)^3}.$$

If $G = 0$, then $d\Pi_1/da = n/8 > 0$; If $s = 0$, then $G = -(1 + a)/4 \Rightarrow d\Pi_1/da > 0$; If $s = 1$, then $G = (1 + a)/4$, and $d\Pi_1/da < 0$.

The bank's incentive to invest in the system is maximal when $d^2\Pi_1/dadG = 0$, which gives $G = -(1 + a)(1 + 2\alpha + \omega)/[4(3 - a)] < 0$ [$d^2\Pi_1/dadG^2 = -(3 - a)\cdot 4n/(1 + a^3) < 0$, which implies that this is a maximum].

Investment as a function of G is illustrated in figure 10.

Figure 10. **Investment when both banks price at the reservation price level**



As to the externalities of private investment, the impact of system improvement on consumer surplus is essential.

The consumer surplus of a bank 1 customer is

$$W_x = G + (2 - i_x) + (1 - s) \cdot \left[a - \frac{(\omega + a)}{2} \right] + s \cdot \left[1 - \frac{(1 + \omega)}{2} \right]$$

and the consumer surplus of a bank 2 customer is

$$W_y = G + (i_x - 1) + s \cdot \left[a - \frac{(\omega + a)}{2} \right] + (1 - s) \cdot \left[1 - \frac{(1 + \omega)}{2} \right].$$

The impact of payment system development on total consumer surplus in the economy equals

$$n \cdot s \cdot \left(\frac{dU_1}{da} \right) + n \cdot (1 - s) \cdot \left(\frac{dU_2}{da} \right) = \frac{n\left(1 + 3a + 3a^2 + a^3 - 32G^2\right)}{\left\{4 \cdot (1 + a)^3\right\}}.$$

Result 2.5.p
Firms invest less than the social optimum in payment system development, at least when G = 0.

Proof:

When G = 0, the impact of the quality of interbank payments on consumer surplus is n/4. The impact of payment system quality on profits is $d\Pi_1/da = d\Pi_2/da = n/8$. The total effect on social welfare equals $n/4 + 2 \cdot n/8 = n/2$, which is much greater than the impact on private profits. Therefore, a private bank has insufficient incentive to invest in the system.

QED

Result 2.5.q
If G = 0 and $\partial^2 a/\partial\Lambda_c\partial\Lambda_z > 0$ ($\partial^2 a/\partial\Lambda_c\partial\Lambda_z < 0$), then the central bank should increase (restrict) its investment in order to encourage the private sector to increase its own investment.

Proof:

When G = 0, private investment is below the social optimum (result 2.5.p).

$d^2\Pi_z/da2 = 2Gn[1 + 2\alpha + 10G + \omega + a(1 + 2\alpha - 2G + \omega)]/(1 + a)^4$, which equals zero when G = 0.

The impact of central bank investment on bank z investment is

$$\frac{d\Lambda_z}{d\Lambda_c} = \frac{\left(\frac{d^2\Pi_z}{da^2}\right)\cdot\left(\frac{\partial a}{\partial \Lambda_z}\right)\left(\frac{\partial a}{\partial \Lambda_c}\right)+\left(\frac{d\Pi_1}{da}\right)\cdot\left(\frac{\partial^2 a}{\partial \Lambda_c \partial \Lambda_z}\right)}{\frac{\partial^2 \Pi_z}{\partial \Lambda_z^2}}$$

$$= \frac{-\left(\frac{d\Pi_1}{da}\right)\cdot\left(\frac{\partial^2 a}{\partial \Lambda_c \partial \Lambda_z}\right)}{\left(\frac{\partial^2 \Pi_z}{\partial \Lambda_z^2}\right)}.$$

Therefore, $d\Lambda_z/d\Lambda_c$ has the same sign as $d^2a/d\Lambda_c d\Lambda_z$.

QED

A few words should also be said about market failures in asymmetric situations. Not surprisingly, a bank with a very small market share invests more than the social optimum in developing the system. (When G = –(1 + a)/4, then s = 0; Then $dU/da + d\Pi_2/da = -n(3\ 2a + 2\alpha + \omega)/[4(1 + a)] < 0$, and investment by bank 1 is socially undesirable.) Both consumers and the rival bank suffer from the investment. The improvement increases the market share of the smaller bank, which, when all the effects are taken into account, reduces the average quality of payment services by increasing the number of interbank payments. Moreover, the rival bank 2 suffers a loss of market share, and the increase in profits of bank 1 is mainly due to the transfer of wealth from the larger to the smaller bank.

2.6 Discussion of the model

2.6.1 Implications of the model

This paper has presented a simple model of duopolistic bank competition. The model describes a Hotelling-type vertically differentiated market where most customers prefer one of the two

banks because of geographic proximity. The emphasis is on payment services rather than lending and borrowing, and especially on banks' incentives to invest in developing the payment system.

Basically, there seem to be two different factors that affect banks' incentives to invest in the payment system:

1) market shares and the degree of concentration in the market
2) interest rate margins and the intensity of competition in financial intermediation and other services.

These factors have rather complicated interaction effects. They largely determine banks' possibilities and incentives to charge fees for using the payment system, which is a key question.

Customers prefer intrabank payments to interbank payments, because intrabank payments are of superior quality. With intrabank payments, there are no delays due to complicated clearing arrangements between the two banks and no exchange of retail payment data between computer systems, which might not be fully compatible. This effect implies that there are economies of scale in the industry. Obviously, the number of intrabank payments is an increasing function of the number of customers who use the same bank. If a bank has only one customer, none of its customer payments are intrabank payments, and if it has a 100 % market share, they are all intrabank payments.

The quality of intrabank payment services is assumed to be exogenous, whereas the quality of interbank payments is determined endogenously. The main issue in this paper concerns the incentives of banks to create an efficient and reliable system for interbank payments. As a rule, banks have incentives to invest in the system in socially non-optimal amounts, especially if they must offer payment services free of charge. In such a case, a small bank may invest excessively in the payment system; otherwise the bank would be unattractive to customers who have to make noncash payments, because there are few agents with whom it would be possible to make intrabank payments. A large bank, by contrast, would have no incentive to improve the interbank system. A well functioning interbank system would weaken its competitive advantage, which is based on its ability to offer a high relative share of intrabank payments.

If the two banks are of the same size, neither of them can increase its market share by improving the system. Thus, banks do not invest in the system, even though customers have pronounced needs for good

interbank payment services. In this case, the central bank should invest in the system.

If banks can charge for payments, their incentives are not as severely distorted. Both banks have incentives to develop the system, because it allows them to charge higher fees. Especially if the two banks are roughly of equal size, there is no serious market failure. The number of interbank payments is high and banks invest substantial amounts in developing the payment system. The central bank should not try to affect the decisions of the private banks. Paradoxically, the degree of price competition depends inversely on the quality of the interbank system: if the system functions poorly, price competition is pronounced.

One of the main conclusions from this model is that banks often have distorted incentives to invest in developing the payment system if they either cannot charge for payment services or they prefer not to do so. If financial market regulation or insufficient interest rate competition results in abnormally wide interest rate margins, attracting more customers often becomes the banks' main objective, because it is the most effective way to increase profits. Thus antitrust policies against collusion regarding loan and deposit services can have beneficial indirect effects on the allocation of resources in the payment system.

A banking monopoly might be optimal for the payment system, but a high degree of concentration would probably not be optimal in terms of allocative efficiency in financial intermediation. The monopoly bank would be able to maintain excessive interest rate margins, which would have serious distorting effects. Thus, there is a welfare tradeoff between efficiency and reliability of the payment system vs socially efficient allocation of financing.

There are three main restrictions incorporated in the model presented in this paper. Without these restrictions, many predictions of the model would probably no longer be valid.

1) In the real world, there are normally more than two banks. Typically, no bank has a dominant market position. If there were three banks, two of them could cooperate, eg via common computer or accounting systems, in order to acquire greater market share at the cost of the third bank. Moreover, with multiple banks it is possible that each bank would have just a minor share of the market.

2) In the real world, the role of payment services as a way to attract customers and deposits is not as simplistic as in the model. For

instance, many households and companies have accounts with numerous banks. Customers might use one bank for payment transactions, and another one for depositing large savings. Thus, it is not obvious that large savings and loan portfolios can be acquired by offering good payment services at low prices.

3) In this model, banks cannot develop their intrabank payment services. Probably large and small banks alike would have incentives to do so, but the large bank might be more interested in developing its internal systems, since most of its customers' payments would be intrabank payments.

Extending the analysis to take these considerations into account might be fruitful.

It is not likely that treating net interest income as an endogenous variable would reverse the results. At least the intuition behind the results is not dependent on the assumption that net interest income is exogenous. The same probably applies, at least to some extent, to the assumed zero price elasticity of demand for payment services.

2.6.2 History of the Finnish payment system

The historical development of the Finnish payment system can be compared with the predictions of the model.

In the late 19th and early 20th century, the most commonplace interbank payment media offered by banks were cheques and so-called postable cashier's drafts[10] (Korpisaari 1930, p. 380–385). Because of their unclear legal status, cheques played a minor role in Finnish payment systems before the legislative reforms in 1920, and they were debt instruments rather than payment media (Kaila 1906, Aaku 1957, p. 145). In the early stages of the postable cashier's draft system, some large banks discriminated against small banks by not accepting instruments issued by the latter. This practice continued until the Bankers' Association was established (Korpisaari 1920, p.

[10] Postable cashier's drafts were used above all in Finland and Sweden. They were debt instruments issued by banks, redeemable upon request, and made payable to a particular payee. The payment was made in the following way. The bank of debtor A sold to its customer a draft payable to creditor B. Debtor A mailed the instrument to payee B. The bank of payee B cashed the instrument, and finally the issuing bank redeemed it with central bank money.

270). This observation is consistent with one of the main results presented in section 2.3, ie that a large bank may try to make it more difficult for the customers of a small bank to make and receive interbank payments.

The savings bank group, consisting of dozens of minor establishments, made several attempts to upgrade payments between member banks. The group launched a kind of giro system already in the 1910s, although this experiment lasted only for a few years (Urbans 1963, p. 392). Group members cooperated in the clearing and issuance of postable cashier's drafts. In 1940 the scope of cooperation was extended, as savings banks began to offer each other's customers all the basic deposit-related services. For instance, it became possible for a customer of savings bank A to make cash withdrawals in the office of savings bank B (Kalliala 1958, p. 60–61). There was no comparable arrangement among commercial banks. On average, commercial banks were much larger than savings banks. As implied by the model, the smallest players in the market were the first to cooperate in respect of payment services.

The current giro system was established during World War II.

For many years, the government had been planning and preparing to establish a postal giro system. The system was finally launched in December 1939, because the Winter War increased the number of retail payments made and received by the government. The government accounted for a disproportionate share of all retail payments. Thus the administrative decision to centralize government payments in the postal savings bank immediately created the critical mass needed to make the system attractive even for private customers. Consequently, the market share of the Postal Savings Bank began to increase in deposit-taking as well. The market share of the Postal Savings Bank was insignificant in the 1930s, but by 1946 it had more than a million deposit customers, which represented a substantial market share; the population of Finland was about 3.8 million (Auer 1964, p. 295). This may be one of the best real world cases corroborating the view that payment services do matter in attracting deposits.

The strengthening role of the Postal Savings Bank caused commercial banks substantial losses of market share. Being forced to react, the commercial banks established their own giro system in 1942. Interestingly, the system was established only between commercial banks; the savings banks did not participate. Because each of the commercial banks operated as an independent agent, and none of them had a dominant market position, one could conclude that the

smallest players in the market were the most interested in launching their own inter-bank giro system.

Due to their close cooperation in payment systems, the savings banks functioned virtually as a single institution. With their nearly 40 % market share, the savings banks did not join the giro system until 1943 (Kuusterä 1995, p. 416). Thus the largest banking group was the last to join the interbank giro system. This observation is consistent with the model presented above. The two giro systems were linked together in 1948.

The model predicts that when there are no fees for using the payment system, a small bank has a strong incentive to develop and maintain a well functioning interbank retail payment system. While the Finnish financial market was tightly regulated during the 1950s, 1960s and 1970s, there was almost no interest rate competition, and payment services were offered free of charge as a marketing tool. None of the Finnish banks had a truly dominant position in the market. Interestingly, no bank attempted to abandon the banks' mutual giro system during those years. Moreover, none of the banks was reluctant to adopt innovations in the mutual exchange of retail payment information, such as physical delivery of magnetic tapes between banks' computer centres in the 1960s and exchange of data via the telephone network in the 1970s. All the banks preferred to participate and to regularly update the technical infrastructure used in interbank payments.

While financial markets were tightly regulated, banks were above all interested in their market shares, possibly because the size of the customer base was essential to profits. Thus it is likely that participating in the giro system was essential to maintaining and increasing the customer base. This interpretation would be consistent with the model: unless a bank has more than a 50 % market share, its market share would decline if it did not participate in the interbank payment system. If there had been a dominant bank in Finland in the past, it might have been reluctant to adapt any innovations that facilitated interbank payments. Eventually, it might even have dropped out of the giro system.

2.6.3 International comparisons of the role of the central bank

One of the key issues of this paper has been the importance of market concentration for the optimal degree of central bank involvement in retail payment systems. Thus it might be interesting to take a closer

look at the situation in different countries. Because the model better describes a giro system than a cheque-based system, the focus in international comparisons should be on countries where cheques play no major role in the payment system.

The following table describes the situation in nine different EU countries. Five member countries (France, Italy, UK, Greece and Ireland) are excluded because their payment systems are cheque-based, and one country (Luxembourg) because of the exceptional nature of its financial industry. The degree of central bank involvement in retail payments is based on a subjective classification, the main source of information being EMI (1996). The degree of market concentration is measured by the three-firm concentration ratio (source of data: EMI).

	The role of the central bank limited	The role of the central bak important
Concentrated banking industry	Finland, Sweden, Denmark, Netherlands	
Intermediate degree of concentration	Belgium	Austria
A fragmented banking industry	Portugal	Spain, Germany

As we see, there seems to be a moderate, negative correlation between the degree of market concentration and the role of the central bank. This would, at least on the surface, be consistent with the conclusions drawn in section 2.3; if the market is concentrated, the central bank should not invest heavily in the system.

Of course, due to the small size of the sample, this evidence has to interpreted with caution. Moreover, it may be due to effects that have little to do with those analysed in this study.

As to countries where cheques are the predominant payment medium, the situation varies greatly. In some of them, the central bank plays a key role in the payment system while the market itself is highly fragmented (United States, Italy) or fairly concentrated (France). In the UK, a cheque system coexists with an extremely limited role for the central bank.

Appendix 1

Result 2.3.f
Let the market share of bank 1 be less than 50 %. Increasing central bank investment reduces bank 1's investment, at least if either or both of the following two conditions hold:

1 $(\partial^2 a/\partial\Lambda_1\partial\Lambda_c) \leq 0$
2 $\Lambda_c = 0.$

Proof:

If the market share is < 50 %, then the investment, Λ_1, is positive.

The bank's investment is determined according to the following optimization condition:

$$\frac{d\Pi_1}{d\Lambda_1} = n \cdot \alpha \cdot \left(\frac{ds}{da}\right) \cdot \left(\frac{\partial a}{\partial \Lambda_1}\right) - 1 = 0.$$

Implicit differentiation gives

$$\frac{d\Lambda_1}{d\Lambda_c} = -\frac{\left(\frac{d^2 s}{da^2}\right) \cdot \left(\frac{\partial a}{\partial \Lambda_c}\right) \cdot \left(\frac{\partial a}{\partial \Lambda_1}\right) + \left(\frac{ds}{da}\right) \cdot \left(\frac{\partial^2 a}{\partial \Lambda_1 \partial \Lambda_c}\right)}{\frac{d^2 \Pi_1}{d\Lambda_1^2}} \cdot n \cdot \alpha$$

where
$ds/da = -G/a^2 > 0$
$d^2s/da^2 = 2G/a^3 < 0$, and
by assumption $da/d\Lambda_c > 0$ and $da/d\Lambda_1 > 0$.

Due to profit maximisation $\left\{d^2\Pi_1 / d\Lambda_1^2\right\} < 0.$
If $(\partial^2 a/\partial\Lambda_1 d\Lambda_c\} \leq 0$, then $d\Lambda_1/d\Lambda_c < 0$ and if $\Lambda_c = 0$, then $\partial a/\partial\Lambda_c = \infty$, which implies

$$\frac{d\Lambda_1}{d\Lambda_c} = \frac{-\infty}{\frac{d^2\Pi_1}{d\Lambda_1^2}} < 0$$

QED

Appendix 2

Result 2.3.g:
If the smaller bank invests less (more) than the social optimum in payment system development, the central bank should restrict (increase) its investment in the system in order to encourage (discourage) private investment, at least if $(d^2a/d\Lambda_1 d\Lambda_c\} \leq 0$.

Proof:

Let z be the bank with the smaller market share.
The optimization condition of the central bank is

$$\frac{d\psi}{d\Lambda_c}=\left(\frac{\partial\psi}{\partial a}\right)\cdot\left[\left(\frac{\partial a}{\partial\Lambda_c}\right)+\left(\frac{\partial a}{\partial\Lambda_z}\right)\cdot\left(\frac{\partial\Lambda_z}{\partial\Lambda_c}\right)\right]-\left(\frac{\partial\Lambda_z}{\partial\Lambda_c}\right)-1=0. \quad (2.3.g.*)$$

If the central bank does not take into account indirect effects, then its optimization condition reduces to

$$\left(\frac{\partial\psi}{\partial a}\right)\cdot\left(\frac{\partial a}{\partial\Lambda_c}\right)-1=0.$$

Because of indirect effects, this optimization policy would make $d\psi/d\Lambda_c$ equal 0 iff

$$\left(\frac{\partial\psi}{\partial a}\right)\cdot\left[\left(\frac{\partial a}{\partial\Lambda_z}\right)\cdot\left(\frac{d\Lambda_z}{d\Lambda_c}\right)\right]-\left(\frac{\partial\Lambda_z}{\partial\Lambda_c}\right)=0.$$

Because $(\partial^2a/\partial\Lambda_1\partial\Lambda_c\} \leq 0$, $(d\Lambda_z/d\Lambda_c) < 0$ (result 2.3.f).
If Λ_z is below the social optimum, then

$$\left(\frac{\partial\psi}{\partial a}\right)\cdot\left(\frac{\partial a}{\partial\Lambda_z}\right)-1>0.$$

If

$$\left(\frac{\partial\psi}{\partial a}\right)\cdot\left(\frac{\partial a}{\partial\Lambda_c}\right)-1=0,$$

then

$$\frac{d\psi}{d\Lambda_c} = \left(\frac{\partial\psi}{\partial a}\right)\cdot\left[\left(\frac{\partial a}{\partial\Lambda_c}\right)+\left(\frac{\partial a}{\partial\Lambda_z}\right)\cdot\left(\frac{\partial\Lambda_z}{\partial\Lambda_c}\right)\right]-\left(\frac{\partial\Lambda_z}{\partial\Lambda_c}\right)-1<0$$

$\Rightarrow (\partial\psi/\partial a)\cdot(\partial a/\partial\Lambda_c) - 1 = 0$ implies that Λ_c would be higher than the social optimum.

Thus, when the central bank takes into account indirect effects, its optimization condition 2.3.g.* implies $(\partial\psi/\partial a)\cdot(\partial a/\partial\Lambda_c) - 1 > 0$.

Therefore, it is optimal to restrict central bank investment. Analogically, if the smaller bank invests more than the social optimum, the central bank should increase its investment.

QED

Appendix 3

Result 2.3.h:
It is optimal for the central bank to invest at least something in the system with any value of G, at least if $\alpha = 1$ and $\partial^2 a/\partial\Lambda_1^2$ is close to zero.

Proof

The impact of central bank investment on social welfare is positive when

$$\frac{d\psi}{d\Lambda_c}=\left(\frac{\partial\psi}{\partial a}\right)\cdot\left[\left(\frac{\partial a}{\partial\Lambda_c}\right)+\left(\frac{\partial a}{\partial\Lambda_1}\right)\cdot\left(\frac{d\Lambda_1}{d\Lambda_c}\right)\right]-\left(\frac{d\Lambda_1}{d\Lambda_c}\right)-1>0$$

which implies

$$\left(\frac{\partial\psi}{\partial a}\right)\cdot\left[1+\frac{\left(\frac{\partial a}{\partial\Lambda_1}\right)\cdot\left(\frac{d\Lambda_1}{d\Lambda_c}\right)}{\frac{\partial a}{\partial\Lambda_c}}\right]-\left\{\frac{\left(\frac{d\Lambda_1}{d\Lambda_c}\right)}{\left(\frac{\partial a}{\partial\Lambda_c}\right)}\right\}-\frac{1}{\left(\frac{\partial a}{\partial\Lambda_c}\right)}>0. \qquad (2.3.h.*)$$

Bank 1 optimizes $\partial\Pi_1/\partial\Lambda_1 - 1 = (ds/da)(\partial a/\partial\Lambda_1) - 1 = 0$. Implicit differentiation gives

$$\frac{d\Lambda_1}{d\Lambda_c}=-n\frac{\left(\frac{d^2s}{da^2}\right)\cdot\left(\frac{\partial a}{\partial\Lambda_c}\right)\cdot\left(\frac{\partial a}{\partial\Lambda_1}\right)+\left(\frac{ds}{da}\right)\cdot\left(\frac{\partial^2 a}{\partial\Lambda_1\partial\Lambda_c}\right)}{\frac{\partial^2\Pi_1}{\partial\Lambda_1^2}}.$$

If central bank investment is close to zero, then

$$\lim_{\Lambda_c\to 0}\left(\frac{\partial a}{\partial\Lambda_c}\right)=\infty.$$

Division by $(\partial a/\partial\Lambda_c)$ gives

$$\frac{\left(\frac{d\Lambda_1}{d\Lambda_c}\right)}{\left(\frac{\partial a}{\partial \Lambda_c}\right)} = -n \frac{\left\{\left[\left(\frac{d^2 s}{da^2}\right)\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right) - \frac{\left[\left(\frac{ds}{da}\right)\cdot\left(\frac{\partial^2 a}{\partial \Lambda_1 \partial \Lambda_c}\right)\right]}{\left(\frac{\partial a}{\partial \Lambda_c}\right)}\right]\right\}}{\frac{\partial^2 \Pi_1}{\partial \Lambda_1^2}} \qquad (2.3.h.^{**})$$

which reduces to

$$\frac{\left(\frac{\partial \Lambda_1}{\partial \Lambda_c}\right)}{\left(\frac{\partial a}{\partial \Lambda_c}\right)} = -n \frac{\left\{\left[\left(\frac{d^2 s}{da^2}\right)\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right)\right]\right\}}{\frac{\partial^2 \Pi_1}{\partial \Lambda_1^2}}$$

$\Rightarrow$ If central bank investment is close to zero, then (2.3.h.*) can be rewritten as

$$\left(\frac{\partial \psi}{\partial a}\right)\cdot\left[1 - \left(\frac{\partial a}{\partial \Lambda_1}\right)\cdot n\left\{\frac{\left[\left(\frac{d^2 s}{da^2}\right)\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right)\right]}{\left(\frac{\partial^2 \Pi_1}{\partial \Lambda_1^2}\right)}\right\}\right] + n\left\{\frac{\left[\left(\frac{d^2 s}{da^2}\right)\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right)\right]}{\left(\frac{\partial^2 \Pi_1}{\partial \Lambda_1^2}\right)}\right\}. \qquad (2.3.h.^{***})$$

Bank 1's optimization implies $da/d\Lambda_1 = 1/(d\Pi_1/da) = -a^2/(G\cdot n\cdot \alpha)$.

On the other hand, $(\partial\psi/\partial a) = n(½ - 2G^2/a^3)$, and

$$\frac{\partial^2 \Pi_1}{\partial \Lambda_1^2} = \frac{\left[\alpha G n \cdot \left(2\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right)^2 - a\cdot\left(\frac{\partial^2 a}{\partial \Lambda_1^2}\right)\right)\right]}{\left(a^3\right)}$$

$\Rightarrow$ (2.3.h.***) can then be rewritten as

$$n \cdot \frac{-2a^6(\alpha-1)+4a^4\alpha G-4\alpha^3\left(\frac{\partial^2 a}{\partial\Lambda_1^2}\right)G^4n^2+a^3G^2\left[8\alpha-8+\alpha^3\left(\frac{\partial^2 a}{\partial\Lambda_1^2}\right)n^2\right]}{-4a^6\alpha+2a^3\alpha^3\left(\frac{\partial^2 a}{\partial\Lambda_1^2}\right)G^2n^2}$$

When $\partial^2 a/\partial\Lambda_1^2 = 0$ and $\alpha = 1$, this reduces to $n(-G/a^2)$; because $G < 0$, this is positive.

It follows that $d\psi/d\Lambda_c > 0$, and therefore the central bank should invest if $G < 0$.

QED

Appendix 4

Result 2.5.g:
In any stable subgame perfect equilibrium, $d\Lambda_z/d\Lambda_c < 0$ if $\partial\Lambda_z/\partial\Lambda_c < 0$ and $d\Lambda_z/d\Lambda_c > 0$ if $\partial\Lambda_z/\partial\Lambda_c > 0$.

Proof:

Both banks optimize their investments according to the reaction function Λ.

The optimal investment depends on three factors:
1) central bank investment
2) rival investment
3) the general popularity of the bank: G for bank 1 and –G for bank 2

The equilibrium of the investment stage is characterized by the following two equations:

$$\Lambda_1 - \Lambda_1(\Lambda_c, \Lambda_2, G) = 0;\ \Lambda_2 - \Lambda_2(\Lambda_c, \Lambda_1, -G) = 0$$

$$|J| = \begin{vmatrix} 1 & -\dfrac{\partial\Lambda_1}{\partial\Lambda_2} \\ -\dfrac{\partial\Lambda_2}{\partial\Lambda_c} & 1 \end{vmatrix} = 1 - \left(\frac{\partial\Lambda_1}{\partial\Lambda_2}\right)\left(\frac{\partial\Lambda_2}{\partial\Lambda_1}\right).$$

Using Cramer's rule, one obtains

$$\frac{d\Lambda_1}{d\Lambda_c} = -\frac{\begin{vmatrix} -\dfrac{\partial\Lambda_1}{\partial\Lambda_c} & -\dfrac{\partial\Lambda_1}{\partial\Lambda_2} \\ -\dfrac{\partial\Lambda_2}{\partial\Lambda_c} & 1 \end{vmatrix}}{|J|} = \frac{\left\{\dfrac{\partial\Lambda_1}{\partial\Lambda_c} + \left(\dfrac{\partial\Lambda_1}{\partial\Lambda_2}\right)\left(\dfrac{\partial\Lambda_2}{\partial\Lambda_c}\right)\right\}}{|J|}.$$

Result 2.5.c and the assumptions concerning the a-function imply $\partial\Lambda_1/\partial\Lambda_2 = \partial\Lambda_2/\partial\Lambda_1$. Unless $|\partial\Lambda_1/\partial\Lambda_2| = |\partial\Lambda_2/\partial\Lambda_1| < 1$, the equilibrium is not stable, since when $|\partial\Lambda_2/\partial\Lambda_1| \geq 1$, it follows that $1 - (\partial\Lambda_1/\partial\Lambda_2)(\partial\Lambda_2/\partial\Lambda_1) \leq 0$.

$\{\partial\Lambda_1/\partial\Lambda_c + (\partial\Lambda_1/\partial\Lambda_2)(\partial\Lambda_2/\partial\Lambda_c)\} = (\partial\Lambda_2/\partial\Lambda_c)[1 + (\partial\Lambda_1/\partial\Lambda_2)]$ is negative (positive) if $(\partial\Lambda_2/\partial\Lambda_c)$ is negative (positive). Consequently, $d\Lambda_1/d\Lambda_c$ is negative (positive) if $\partial\Lambda_1/\partial\Lambda_c$ is negative (positive).

QED

Appendix 5

Result 2.5.i
It is optimal for the central bank to invest nothing in developing the system if $|G| \geq 3a/(2\cdot\sqrt{[2 + 1/a]})$. If G = 0, it is optimal to invest.

Proof:

$$\frac{da}{d\Lambda_c} = \frac{\partial a}{\partial \Lambda_c} + \left(\frac{\partial a}{\partial \Lambda_1}\right)\cdot\left(\frac{d\Lambda_1}{d\Lambda_c}\right) + \left(\frac{\partial a}{\partial \Lambda_2}\right)\cdot\left(\frac{d\Lambda_2}{d\Lambda_c}\right)$$
$$= \frac{\partial a}{\partial \Lambda_c} + 2\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right)\cdot\left(\frac{d\Lambda_1}{d\Lambda_c}\right)$$

where

$$\frac{d\Lambda_1}{d\Lambda_c} = \frac{\left\{\frac{\partial \Lambda_1}{\partial \Lambda_c}\left[1 + \left(\frac{\partial \Lambda_1}{\partial \Lambda_2}\right)\right]\right\}}{\left\{1 - \left(\frac{\partial \Lambda_1}{\partial \Lambda_2}\right)^2\right\}} = \frac{\left(\frac{\partial \Lambda_1}{\partial \Lambda_c}\right)}{\left\{1 - \left(\frac{\partial \Lambda_1}{\partial \Lambda_2}\right)\right\}}$$

(see appendix 4).

The effect on welfare is

$$\frac{d\psi}{d\Lambda_c} = \frac{\partial \psi}{\partial a}\left[\frac{da}{d\Lambda_c}\right] - 1 - 2\cdot\frac{d\Lambda_1}{d\Lambda_c}$$
$$= \frac{\partial \psi}{\partial a}\left[\frac{\partial a}{\partial \Lambda_c} + \frac{2\cdot\left(\frac{\partial a}{\partial \Lambda_1}\right)\cdot\left(\frac{\partial \Lambda_1}{\partial \Lambda_c}\right)}{\left\{1 - \left(\frac{\partial \Lambda_1}{\partial \Lambda_2}\right)\right\}}\right] - 1 - \frac{2\left(\frac{\partial \Lambda_1}{\partial \Lambda_c}\right)}{\left\{1 - \left(\frac{\partial \Lambda_1}{\partial \Lambda_2}\right)\right\}}. \qquad (2.5.i.*)$$

When $\Lambda_c \approx 0$, $\partial\Lambda_1/\partial\Lambda_c > 0$ (result 2.5.h).

When $|G| = 3a/(2\cdot\sqrt{[2 + 1/a]})$,
$d\psi/d\Lambda_c = -1 - 2(\partial\Lambda_1/\partial\Lambda_c)/\{1 - (\partial\Lambda_1/\partial\Lambda_2)\} < 0$.

If $|G| > 3a/(2\cdot\sqrt{[2 + 1/a]})$, then the impact of central bank investment on welfare becomes even more negative, because

$d\psi/da[\partial a/\partial \Lambda_c + 2\cdot(\partial a/\partial \Lambda_1\}\cdot(\partial \Lambda_1/\partial \Lambda_c)/\{1 - (\partial \Lambda_1/\partial \Lambda_2)\}] < 0$

$\Rightarrow$ Increasing central bank investment above zero is not optimal if $|G| \geq 3a/(2\cdot\sqrt{[2 + 1/a]}$.

If $G = 0$, then according to result 2.5.e, $(d\psi/da)\cdot(\partial a/\partial \Lambda_1) - 1 = 0$, and (2.5.i.*) reduces to $\partial\psi/\partial a[\partial a/\partial \Lambda_c] - 1$.

When $\Lambda_c = 0$, $\partial a/\partial \Lambda_c = \infty$.

Because $\partial\psi/\partial a > 0$, $\partial\psi/\partial a[\partial a/\partial \Lambda_c] - 1 > 0$.

Thus it cannot be optimal not to invest when $G = 0$.

QED

Appendix 6

Result 2.5.n
If investment by bank z (z = 1, 2) is greater than zero, the impact of central bank investment on private investment, Λ_z, is negative, at least if either $\Lambda_c \approx 0$ or $\partial^2 a/\partial\Lambda_z\partial\Lambda_c \leq 0$.

Proof:

The f.o.c. of bank 1 is $d\Pi_1/d\Lambda_c = 1$.

Implicit differentiation gives

$$\frac{d\Lambda_1}{d\Lambda_c} = -\frac{\left[\frac{\partial^2\Pi_1}{\partial\Lambda_1\partial\Lambda_c}\right]}{\left[\frac{\partial^2\Pi_1}{\partial\Lambda_1^2}\right]} = \frac{2\left(\frac{\partial a}{\partial\Lambda_c}\right)\cdot\left(\frac{\partial a}{\partial\Lambda_1}\right) - a\left(\frac{\partial^2 a}{\partial\Lambda_1\partial\Lambda_c}\right)}{-2\left(\frac{\partial a}{\partial\Lambda_1}\right)^2 + a\left(\frac{\partial^2 a}{\partial\Lambda_1^2}\right)}.$$

When either $\Lambda_c = 0$ (and consequently $\partial a/\partial\Lambda_c = \infty$) or $\partial^2 a/\partial\Lambda_1\partial\Lambda_2 \leq 0$, $d\Lambda_1/d\partial\Lambda_c < 0$.

Bank 2 f.o.c is $d\Pi_2/d\Lambda_2 - 1 = 0$.

$$\frac{d\Lambda_2}{d\Lambda_c} = \frac{\left[\frac{\partial^2\Pi_2}{\partial\Lambda_2\partial\Lambda_c}\right]}{\left[\frac{\partial^2\Pi_2}{\partial\Lambda_2^2}\right]}$$

$$= \frac{2(\alpha - 2G)^2\cdot\left(\frac{\partial a}{\partial\Lambda_2}\right)\left(\frac{\partial a}{\partial\Lambda_c}\right) + a\left[a^2 - (\alpha - 2G)^2\right]\left(\frac{\partial^2 a}{\partial\Lambda_c\partial\Lambda_2}\right)}{-2(\alpha - 2G)^2\cdot\left(\frac{\partial a}{\partial\Lambda_2}\right) - a\left\{a^2 - (\alpha - 2G)^2\left(\frac{\partial^2 a}{\partial\Lambda_2^2}\right)\right\}}. \quad (2.5.n.*)$$

Because $G < 0$, $s = (3a - \alpha + 2G)/(4a) < 1/2$, which implies $3a - \alpha + 2G < 2a$, which in turn implies $2G - \alpha < -a$. Because $-a < 0$, $2G - \alpha < 0$.

It follows that $|2G - \alpha| > |a| \Rightarrow + a^2 - (\alpha - 2G)^2 < 0$.
This, in turn, implies

$$-2(\alpha - 2G)^2 \cdot \left(\frac{\partial a}{\partial \Lambda_2}\right)^2 - a\left\{a^2 - (\alpha - 2G)^2\right\}\left(\frac{\partial^2 a}{\partial \Lambda_2^2}\right)$$

$d\Lambda_2/d\Lambda_c$ is negative when the numerator of 5n.* is positive. This is true at least when either $\Lambda_c = 0$ or $\partial^2 a/\partial\Lambda_2\partial\Lambda_c \leq 0$ or both.

QED

3 Discriminatory patent protection: Two Extensions of the Aoki–Prusa model

Patent systems in different parts of the world have sometimes favoured domestic applicants. This discrimination against foreigners has taken place in both explicite provisions and in tacit administrative practice. In the US, the legal process for a patent holder whose patent rights are violated has been more complicated in the case of foreign patent holders (Aoki and Prusa, 1993). Kotabe (1992) used a purely statistical approach and found evidence for discriminatory practices in several countries. Japanese authorities discriminate against foreign applicants with longer pendency periods. In the United States, United Kingdom and Germany, it appears that foreigners are discriminated against via lower patent grant ratios. Some kind of discrimination has been practiced by the EU as well (Schwartz, 1991).

In principle such practices are now banned among WTO member countries. Non-discriminatory treatment is guaranteed by the Trade-Related Aspects of Intellectual Property Rights Agreement (TRIPS-agreement), which was negotiated in 1994 during the GATT Uruguay Round. The contract came into force in January 1995 but with lengthy transition periods ranging from one to eleven years, depending on the country. Thus, if such an international agreement contributes to welfare, or has any other consequences, the effects should already be materializing. Hence the potential benefits of non-discriminatory patent laws is a highly topical issue.

Interestingly, it is relatively difficult to find much theoretical work on this apparently relevant topic. The model presented by Aoki–Prusa (1993) is one of the very few contributions in this area. Their work has not inspired much further research, although Adams (1998) extended the model in a relatively short paper focusing on the competition between technology leaders in the North and followers in the South. Moreover, Taylor (1994) presented a model on discrimination that is based on the location of the R&D, rather than on the home country of the patentee.

Aoki–Prusa model

One of the few theoretical contributions that analyse the impact of discriminatory patent protection was presented by Aoki and Prusa (1993). Motivated mainly by the discriminatory practice applied in the US, they analyse the impact of such discriminatory patent policies on the R&D incentives of firms, especially those that may have been granted privileges by their national governments.

In the model presented in their paper, Aoki–Prusa assume that things happen in the following order:

1) Both a domestic firm and its foreign rival invest in R&D. They are working on the same product or process idea.
2) If one manages to develop the technology, it tries to patent it. If both firms have the technology, they both try to patent it.
3) Public authorities grant patent rights.
4) The product is produced and sold by the patentee.

If only one of the firms files the application, it becomes the patentee. If both firms do, they can face either discriminatory or non-discriminatory legislation. In the discriminatory case, the domestic firm gains full monopoly rights. In the non-discriminatory case, both firms gain the right to use the invention, and they both earn some profits as a result, although the sum of their profits would be lower than the monopoly profits earned by the domestic firm in the discriminatory case. One of the main conclusions of the paper is that if the new invention does not replace an old technology, the discriminatory policy will necessarily induce the domestic firm to spend more on innovative activities than would non-discriminatory legislation.

In addition to their basic model, Aoki and Prusa (1993) analyse a situation where the domestic firm competes against a relatively inefficient foreign firm. They show that discrimination can reduce domestic R&D if there are older products that would be replaced by the eventual invention. The threat of the foreign rival capturing a very large market share almost disappears if the rival is both inefficient and discriminated against, and the favoured domestic firm can count on relatively high profits in the future even if it relies on its old technologies. Hence, discrimination would to an important extent eliminate the incentive to carry out costly R&D.

The following analysis is not related to this extension of the basic model. Instead, it extends the basic model of Aoki–Prusa in two different ways. First, it is demonstrated that if there are two countries

where the product can be marketed, simultaneous discrimination by both governments may be irrelevant to firms' R&D decisions.

Then, it will be demonstrated that if there is only one government that practices discrimination, this policy may actually provide a disincentive for the domestic firm. This result is obtained by slightly altering the assumptions of the Aoki–Prusa model. The result is valid even if we do not assume that there would be any older inventions that would become worthless because of the new invention.

3.1 The first extension: A two-country world

3.1.1 Assumptions

It is realistic to assume that if national legislation discriminates against foreign firms, the impact is limited to the domestic territory. Abroad, the firms may be treated equally, or possibly the foreign government gives its domestic firm certain privileges. However, Aoki–Prusa assume that there is no potential market for the product outside the country where discriminatory policies are applied. Now, we shall see how the conclusions are affected if we assume a symmetric situation where both firms have a home country where the product can be sold.

There are two firms, 1 and 2, trying to invent a new product. In addition, there are two countries, 1 and 2, where the product can be sold. Firm i is owned by nationals of the country i. Things happen in the following order.

1) The governments decide simultaneously on the degree of discrimination in patent legislation. The governments do not cooperate. Firms can immediately observe these decisions.
2) Firms decide on their R&D expenditures simultaneously, without collusion.
3) The innovator(s) file(s) patent applications in both countries. The patentees are chosen.

The probability that firm i invents is $p_i = p(r_i)$, where r_i is the R&D effort of firm i. If a firm has invented something, it will try to patent the invention. The value of the patent for country i is α_i. The value of the patent is exogenously given. A firm with no patents earns no return on its R&D investment.

If only one of the firms innovates, it becomes the patentee in both countries. As the holder of legal monopoly rights in both countries, the firm will earn revenue worth $\alpha_1 + \alpha_2$.

If neither of the firms invents, there are no patents, and the firm earns no revenue. Hence R&D investments are lost.

If both firms innovate, the two governments will choose the patentee independently of the other. The probability that the domestic firm of country i wins the patent contest in its home country is z_i. If the government applies non-discriminatory policies, the two firms have equal opportunities to get the patent ($z_i = ½$). If the government practices maximal discrimination, the domestic firm will automatically get the patent if it has innovated ($z_i = 1$).

Unlike in the original model of Aoki and Prusa, it is now assumed that the government will never divide patent rights between the two firms. So far as the author knows, such division of patent rights is not practiced anywhere. It may be interesting to explore why this is the case, but this is beyond the scope of the present study.

With these assumptions, firm i's expected profit is the difference between expected revenue from the invention and R&D costs, ie

$$\pi_i = \alpha_i \cdot p_i(r_i) \cdot \left[1 - (1 - z_i) \cdot p_j\right] + \alpha_j \cdot p_i(r_i) \cdot \left[1 - z_j \cdot p_j\right] - r_i. \qquad (3.1.i)$$

As to the functional form of the p-function, it is assumed that both firms have identical p-functions. The functions are characterized by the following conditions:

If $r_i = 0$, then $dp_i/dr_i = \infty$,

$p_i < 1$ with any finite value of r_i.

These two assumptions guarantee that both firms always invest in R&D, implying that they both have a probability to innovate that is positive but less than 1.

If either of the two firms gets the patent, the new product will be available to consumers. Consumers have no preferences as to the supplier. The price consumers have to pay for the product does not depend on who is the patent holder. If the good becomes available in the shops, the overall increase in consumer surplus in country i is Γ_i, when compared to the situation without the good.

3.1.2 Solving the model

The first result has strong analogies with the first main result presented by Aoki and Prusa: in the simplest case, discrimination encourages domestic R&D and increases domestic profits, and there is no reason not to discriminate.

Result 3.1.a
If both governments try to maximize the expected value of domestic profits, and if the governments have not agreed on forbidding discriminatory patent policies, both governments practice maximally discriminating policies ($z_1 = z_2 = 1$).

Proof:

Appendix 1

This result is not difficult to understand intuitively. The discrimination simply decreases the risk that the domestic R&D effort would turn out to be useless, and it has no other effects.

One of the main implications of the proof of result 3.1.a is that domestic R&D would be maximized by maximal domestic discrimination. It was found that $dr_i/dz_i > 0$. Hence, if the governments tried to maximize domestic R&D instead of domestic corporate profits, as Aoki and Prusa assumed, both governments would still apply maximal discrimination. This result has very strong analogies with the main finding of Aoki and Prusa.

The same equilibrium would prevail if both governments tried to maximise social welfare instead of domestic profits. In that case, the government's objective function would include consumers' utility. Both foreign and domestic R&D efforts might be affected by changes in patent legislation. This, in turn, implies that patent policies may be relevant to consumers' possibilities of finding the good in the shops. Nevertheless, if the analysis is restricted to symmetric cases, the optimal stategy for both governments will always be to discriminate maximally against the foreign firm.

Result 3.1.b
In symmetric cases (ie where $\alpha_1 = \alpha_2$, $z_1 = z_2$, $\Gamma_1 = \Gamma_2$) the following holds. If both governments try to maximize social welfare, defined as the sum of domestic corporate profits and domestic consumer

surplus, the only possible equilibrium is characterized by maximal discrimination by both governments; $z_1 = z_2 = 1$.

Proof:

Appendix 2.

This result is largely due to the fact that in any symmetric case, consumer surplus is marginally invariant to changes in the degree of discrimination. If the firm is discriminated against in the foreign country, its incentives to invest in R&D are weakened. However, the incentives of the foreign company are strengthened. These two effects offset each other, and consumers' possibilities of finding the product in the shops are unchanged. Hence whenever $z_1 = z_2$, a government cannot affect consumer surplus by marginally adjusting its z-parameter. Thus the only factor that matters in government policies is the impact of discrimination on domestic profits. Consumer surplus cannot be affected.

The following result may seem even more surprising.

Result 3.1.c
If the analysis is restricted to symmetric cases, as defined above, the value of $z = z_i = z_j$ does not affect firms' expected profits or R&D efforts.

Proof:

According to 3.1.i firm i's expected profit is:

$$\pi_i = p_1 \cdot \alpha \cdot \{1 - (1 - z) \cdot p_j\} + p_i \cdot \alpha(1 - z \cdot p_j) - r_i$$
$$= 2 \cdot p_i \alpha - p_j p_i \alpha - r_i$$

Therefore, $\partial\pi_i/\partial z \equiv 0$, and firms have no reason to adapt their R&D efforts according to discriminatory patent policies, provided both governments discriminate equally strongly.

QED

Corollary of the result 3.1.c
If the analysis is restricted to symmetric cases (ie where $z_1 = z_2$, $\alpha_1 = \alpha_2$, $\Gamma_1 = \Gamma_2$ and $r_1 = r_2$), then R&D efforts, expected profits, expected consumer surplus and consequently expected social welfare do not depend on the degree of discrimination.

Therefore, in such a symmetric case, an international agreement banning the use of discriminatory patent policies would not benefit either of the countries. This result is an obvious implication of result 3.1.c. In result 3.1.c, it was demonstrated that as long as the two z-parameters are equal, firms' expected profits and R&D efforts are invariant to the degree of discrimination. Because the R&D investments do not depend on discrimination, the likelihood that the product will be available for consumers does not depend on patent policies.

Traditional models concerning protectionism assume that the domestic firm enjoys tariff protection against foreign competition in the end-product market. Standard theoretical analysis implies that such policies probably reduce globally welfare. This theoretical result has rather strong policy implications; the fact that governments' economic advisors are normally well informed about traditional trade theories may be one of the main reasons why international free trade agreements are as commonplace as they are. But, as we have seen, it is far from obvious that the result would have direct implications in the case of discriminatory patent policies. In fact, very few assumptions applied in the standard analysis of tariff protection satisfactorily describe a patent contest. International agreements that oblige countries to guarantee equal intellectual property rights to foreigners may be both useless and harmless.

If the firms are assumed to be risk-averse expected utility maximisers, even more curious results could emerge. It is quite possible that both of them would prefer a relatively high degree of mutual discrimination. In the above model, excessively risk-averse firms with strongly concave utility functions would above all want to minimise the risk that the rival wins the patent in both countries. This would be relatively unlikely if the firm could count on getting the patent at least in its home country. By contrast, the possibility to get the patent in both countries would not be a major incentive for such a firm.

One has to remember one key limitation of this result. In the model, it was assumed that both firms are more or less equally competent in technology issues. Discriminatory patent policies are likely to reduce welfare and economic growth – or at least R&D

efforts – if the government impedes foreign firms from obtaining patents even in the complete absence of domestic firms that could possibility develop the same technology. If, for instance, an LDC country with almost no R&D were to deny intellectual property rights to foreigners, this would be a disincentive to foreign R&D efforts. However, it would not encourage domestic inventors, because there would not be any.

In addition, result 3.1.c has another major limitation. It is based on the assumption that the two countries are of equal size. If this assumption is relaxed, it is no longer obvious that mutual discrimination would not have any effects.

Result 3.1.d
If either of the two countries is bigger than the other as a market area ($\alpha_1 \neq \alpha_2$), and if both governments try to maximize domestic R&D, an international agreement forbidding discrimination in patent policies would increase R&D in the smaller country but reduce it in the big country.

Proof:

Appendix 3.

The result is fairly intuitive. Both firms are assumed to be equally efficient producers and users of technological know-how. However, the government of the bigger country can practice a much more efficient discriminatory policy simply by controlling access to a much larger market area. To take an extreme example, a Luxembourg-based firm competing against a German rival would not benefit much if its national government favoured it in the patent contest, because nearly all the potential customers are in Germany. Therefore, from the point of view of a small country, it might be useful to include a ban on discriminatory patent policies in international free trade agreements, whereas big countries do not necessarily have such interests.

3.2 The second extension: What if there will be a patentee in any case?

3.2.1 Assumptions

Unlike in the first extension, but exactly as in the original model of Aoki and Prusa, it is now assumed that there is only one country where the product can be sold. Factors related to culture, legislation and climate could make it impossible to sell the product in other countries. For instance, a new heating system for residential buildings would have no demand in Singapore, even though the product might be developed in the tropics. Established technical standards in certain countries might be incompatible with a new invention: better GSM phones are of no use if the local mobile phone network is based on an entirely different technical standard.

Here it is assumed that one of the two companies is owned by citizens of the home country, whereas its rival is owned by foreigners. As in the original model of Aoki and Prusa, the government is assumed to aim at maximal domestic R&D. In the following, it will be demonstrated that even in the case of one country, the main result of the basic model of Aoki and Prusa is sensitive to changes in the assumptions.

It is assumed that R&D efforts cannot be totally useless. There will be some innovative output, and a patent will be granted in any case, either to the domestic firm or to the foreign one. Both firms will file a patent application for its research results, and one of the two firms will get the patent.

Investing in R&D is useful for two different reasons. First, R&D increases the likelihood that the firm will get the patent. The probability that the domestic firm (1) will get the patent is denoted f,

$$f = \left\{ \frac{r_1}{(r_1 + r_2)} \right\}^{1-z} . \qquad (3.2.i)$$

Now r_i is the whole stock of relevant technological knowledge of firm i. This stock consists of R&D carried out in the patent contest analysed in this model and knowledge acquired from other sources, including publicly available knowledge and knowledge acquired in

past, completed projects of the firm. $r_i = v_i + c_i$, where v_i is the exogenous stock of relevant knowledge and c_i is the R&D expenditure of the firm.[11]

Obviously the probability that the foreign firm (2), gets the patent is 1–f. The parameter z measures the degree of discrimination. The lowest possible value of z is zero. If z = 0, there is no discrimination; the probability that firm i wins the patent is $r_i/(r_i + r_j)$, irrespective of whether i = 1 or i = 2. The parameter z cannot be larger than 1; in the extreme case where z = 1, discrimination is excessive, and the domestic firm always gets the patent.

Secondly, R&D increases the value of the invention, and makes the eventual patent more valuable. If R&D investment has been low, the patent will be almost valueless. If firm i gets the patent, it will earn revenue of α_i, $\alpha_i = \alpha(r_i)$. The value of the eventual patent depends only on the firm's own R&D expenditure. The function has the properties $d\alpha_i/dr_i > 0$ and $d^2\alpha_i / dr_i^2 < 0$. No other assumptions concerning the functional form of α are made. Denote $d\alpha_i/dr_i$ as α_i'.

Events happen in the following order. First, the domestic government sets the value of z and reveals its decision to both firms. Secondly the two firms simultaneously choose their R&D efforts. Then the patentee is chosen, and the product is commercialized by the patentee. The expected profit of the domestic firm is

$$\pi_1 = \alpha_1 \cdot f - c_1 \qquad (3.2.ii)$$

and the foreign firm's expected profit is

$$\pi_2 = \alpha_2 \cdot (1 - f) - c_2. \qquad (3.2.iii)$$

Again, it is assumed that if the government does not practice discrimination (z = 0), the two firms end up in a symmetric equilibrium where $r_1 = r_2$, and consequently both firms have equal possibilities to get the patent. In addition, this symmetry implies that for both firms' the probability that it will get the patent depends symmetrically on its R&D efforts. (When z = 0, then $\partial f/\partial r_1 = -\partial f/\partial r_2$)

[11] If it were assumed that the total cost of participating in the patent contest equals r, the R&D cost would often exceed the expected revenue. In the symmetric equilibrium analysed in the following firms would actually minimize losses instead of maximizing profit. Thus, not participating (r=0) would be a better alternative. This problem is avoided if r can contain information the firm can use in this patent contest without paying for it.

This Nash equilibrium is assumed to be stable. Both firms are again assumed to be risk neutral.

3.2.2 An analytical result

This section focuses on a result that is valid in very specific cases only. It is of interest because it demonstrates what kinds of effects may arise if discrimination is practised.

Result 3.2.a
Let the situation be symmetric (ie $r_1 = r_2 = r$, $\alpha_1 = \alpha_2$, $f = ½$ and $z = 0$). $dr_1/dz < 0$ at least in cases where $\alpha' \approx 0$. $dr_1/dz > 0$ at least in cases where α' has the highest value it can have in a meaningful, symmetric equilibrium.

Proof: Appendix 4.

QED

This mathematical result has the following interpretation. If R&D matters mainly because it increases the value of the patent, then discrimination encourages domestic R&D. This is easy to this understand intuitively. Discrimination would simply reduce the risk associated with the project, ie the risk that the foreign rival would get the patent. This could be called the risk effect. This result has very strong analogies with the main result of the basic model of Aoki and Prusa.

If R&D is useful mainly because it helps to win the patent contest ($\alpha' \approx 0$), the situation is completely different. In such a case, discrimination would actually *reduce* domestic R&D. This outcome could be called the competition effect. If the firm is favoured by its domestic government, it does not have to work hard in order to get the patent. The possibility that discrimination might free the domestic firm from competitive pressures and thereby enable a reduction in R&D efforts does not appear in the original model of Aoki and Prusa.

3.2.3 Simulation results

One should note that the result presented above, in section 3.2.2, has a very important restriction: it was proved to be valid for a non-discriminatory, symmetric equilibrium only. Even if in the non-discrimination equilibrium a slight marginal discrimination would either increase or decrease domestic R&D, it may be possible that with higher degrees of discrimination ($z >> 0$), this effect would be reversed.

Relaxing the restriction that $z = 0$ would complicate the mathematical analysis significantly. In order to be able to draw more general conclusions the situation is now analysed using simulations. Here it is assumed that the value of the invention follows the functional form $\alpha_i = a \cdot Ln(1+r_i)$, where a is an exogenous parameter common to both firms, and α_i is the value of the invention for firm i. The impact of z on the possibility of getting the patent is assumed to follow the functional form presented in formula 3.2.i.

The following two tables present the simulation results obtained with systematic experiments with different values of a and z. The figures present the Nash equilibrium values of R&D outlays.

The algorithm is fairly simple. For each combination of parameters, the equilibrium is found in the following way. First, it is assumed that $r_2 = 0$. Then experiments are carried out to find the value of r_1 that maximizes π_1. Then, assuming that r_1 takes the value obtained in the first experiment, the value of r_2 is found in a similar way. Then, the same procedure is applied again to r_1, then to r_2, and so on, until the simulations converge.

Table 1. **Simulation results concerning R&D expenditure of the domestic firm as a function of discrimination. Maximal domestic R&D is in bold. (x = no equilibrium found)**

	a=5.	a=10	a=20	a=40	a=80	a=160	a=320	a=640	a=1280
z=0.00	4.018	10.729	26.136	60.938	138.471	**309.236**	**681.867**	**1488.807**	**3225.089**
z=0.05	**4.371**	11.051	26.526	61.398	**138.887**	309.221	680.327	1482.936	3208.158
z=0.10	x.xxx	11.290	26.789	**61.615**	138.821	308.208	676.696	1472.692	3181.988
z=0.15	x.xxx	11.446	**26.917**	61.582	138.249	306.147	670.873	1457.827	3146.072
z=0.20	4.000	**11.507**	26.903	61.283	137.143	302.984	662.741	1438.089	3099.891
z=0.25	4.000	11.441	26.736	60.702	135.476	298.657	652.170	1413.223	3042.881
z=0.30	4.000	xx.xxx	26.393	59.818	133.211	293.105	639.035	1382.950	2974.477
z=0.35	4.000	xx.xxx	25.835	58.601	130.306	286.250	623.191	1346.985	2894.062
z=0.40	4.000	9.000	xx.xxx	57.005	126.710	278.010	604.477	1305.025	2801.026
z=0.45	4.000	9.000	xx.xxx	54.946	122.344	268.275	582.713	1256.715	2694.662
z=0.50	4.000	9.000	xx.xxx	xx.xxx	117.088	256.905	557.680	1201.656	2574.235
z=0.55	4.000	9.000	19.000	xx.xxx	110.685	243.669	529.085	1139.393	2438.904
z=0.60	4.000	9.000	19.000	xx.xxx	xxx.xxx	228.129	496.446	1069.232	2287.576
z=0.65	4.000	9.000	19.000	39.000	xxx.xxx	xxx.xxx	458.795	990.133	2118.715
z=0.70	4.000	9.000	19.000	39.000	79.000	xxx.xxx	xxx.xxx	xxx.xxx	1929.362
z=0.75	4.000	9.000	19.000	39.000	79.000	159.001	xxx.xxx	xxx.xxx	xxx.xxx
z=0.80	4.000	9.000	19.000	39.000	79.000	159.001	319.001	638.995	xxx.xxx
z=0.85	4.000	9.000	19.000	39.000	79.000	159.001	319.001	638.995	1278.991
z=0.90	4.000	9.000	19.000	39.000	79.000	159.001	319.001	638.995	1278.991
z=0.95	4.000	9.000	19.000	39.000	79.000	159.001	319.001	638.995	1278.991
z=1.00	4.000	9.000	19.000	39.000	79.000	159.001	319.001	638.995	1278.991

Table 2. **Simulation results concerning R&D expenditure of the foreign firm as a function of discrimination. (x = no equilibrium found)**

	a=5.	a=10	a=20	a=40	a=80	a=160	a=320	a=640	a=1280
z=0.00.	4.018	10.729	26.136	60.939	138.471	309.236	681.867	1488.804	3225.091
z=0.05	3.435	10.061	24.966	58.686	133.964	300.049	662.961	1449.689	3143.959
z=0.10	x.xxx	9.280	23.623	56.110	128.817	289.567	641.393	1405.074	3051.422
z=0.15	0.000	8.366	22.094	53.195	123.002	277.727	617.037	1354.678	2946.883
z=0.20	0.000	7.278	20.364	49.921	116.485	264.464	589.752	1298.233	2829.763
z=0.25	0.000	5.895	18.411	46.266	109.229	249.711	559.402	1235.435	2699.443
z=0.30	0.000	x.xxx	16.196	42.201	101.194	233.391	525.839	1165.986	2555.289
z=0.35	0.000	x.xxx	13.639	37.687	92.331	215.424	488.908	1089.566	2396.623
z=0.40	0.000	0.000	xx.xxx	32.660	82.580	195.719	448.441	1005.848	2222.809
z=0.45	0.000	0.000	xx.xxx	26.998	71.860	174.174	404.270	914.498	2033.153
z=0.50	0.000	0.000	xx.xxx	xx.xxx	60.025	150.646	356.191	815.170	1826.987
z=0.55	0.000	0.000	0.000	xx.xxx	46.762	124.920	303.934	707.446	1603.557
z=0.60	0.000	0.000	0.000	xx.xxx	xx.xxx	96.524	247.115	590.856	1362.078
z=0.65	0.000	0.000	0.000	0.000	xx.xxx	xx.xxx	184.840	464.508	1101.487
z=0.70	0.000	0.000	0.000	0.000	0.000	xx.xxx	xxx.xxx	xxx.xxx	819.753
z=0.75	0.000	0.000	0.000	0.000	0.000	0.000	xxx.xxx	xxx.xxx	xxx.xxx
z=0.80	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	xxx.xxx
z=0.85	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
z=0.90.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
z=0.95	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
z=1.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

These simulation results reveal at least the following regularities.

1) Maximal discrimination (z = 1) does not maximize domestic R&D under any circumstances. On the contrary, domestic R&D is maximized either by nondiscrimination or relatively mild discrimination.
2) If the value of the invention is relatively high (a >> 0), domestic R&D is maximized with nondiscrimination.
3) If the value of the invention is relatively low (a is low), domestic R&D is maximized via imperfect discrimination.
4) In no case does more discrimination encourage the foreign firm to increase its R&D effort. This in turn implies that a domestic government that practices discrimination always hampers technological development in the foreign country.

Hence, at least in the case of this specific functional form, a government interested in maximizing domestic R&D should never practice maximal discrimination. The difference between this outcome and the main result of the basic model by Aoki & Prusa is quite pronounced.

3.3 Conclusions

In real life, governments have practised discriminatory patent policies; it may be easier for a domestic firm to protect its inventions with patents. Aoki and Prusa have presented a theoretical analysis of the impact of such policies on firms' R&D efforts. The analysis presented above extends the basic model of Aoki and Prusa. At a very general level one can say that the most important finding may be that discriminatory patent policies might have surprising effects, and no simplistic generalizations concerning the dangers and benefits of such policies seem to be robust.

To be more specific, it was demonstrated that if two countries discriminate against each other's firms in order to favour domestic companies, the aggregate effect may be nil. At the firm level, being favoured in the home country and discriminated against in the foreign country, may have offsetting effects. Thus, if one takes a global approach, discriminatory patent policies may be harmless.

Secondly, it was demonstrated that discrimination may either encourage or discourage domestic R&D, because R&D efforts may be aimed at slightly different objectives. If a marginal increase in R&D would be useful mainly because it improves the possibilities of getting the patent, discrimination might actually discourage the domestic firm in its R&D efforts. This discrimination is a good substitute for costly R&D efforts in the patent contest, and a favoured domestic firm does not have to put much effort into the project. If, instead, R&D mainly increases the value of the invention, discrimination mainly makes it less likely that the invention will be lost and thereby increases the incentives for R&D by reducing the associated risk.

Appendix 1

Result 3.1.a
If both governments try to maximize the expected value of domestic profits, and if the governments do not agree on forbidding discriminatory patent policies, both governments practice maximally discriminating policies ($z_1 = z_2 = 1$).

Proof:

According to 3.1.i the expected profit of firm i (i = 1 or 2) is $\pi_i = p_i(r_i)\cdot\alpha_i\cdot\{1-[1-z_i]\cdot p_j(r_j)\} + p_i(r_i)\cdot\alpha_j[1-z_j\cdot p_j(r_j)] - r_i$.

The first order conditions for the firms are

$$\frac{\partial\pi_i}{\partial r_i} = \left(\frac{\partial p_i}{\partial r_i}\right)\cdot\alpha_i\cdot\left\{1-(1-z_i)\cdot p_j\right\} + \left(\frac{\partial p_i}{\partial r_i}\right)\cdot\alpha_j\left(1-z_j\cdot p_j\right) - 1 = 0$$

$$\frac{\partial\pi_j}{\partial r_j} = \left(\frac{\partial p_j}{\partial rj}\right)\cdot\alpha_j\cdot\left\{1-(1-z_j)\cdot p_i\right\} + \left(\frac{\partial p_j}{\partial r_j}\right)\cdot\alpha_i\left(1-z_i\cdot p_i\right) - 1 = 0. \qquad (3.1.a.*)$$

The Jacobian of this system is

$$\left(\frac{\partial^2\pi_i}{\partial r_i^2}\right)\cdot\left(\frac{\partial^2\pi_j}{\partial r_j^2}\right) - \left(\frac{\partial^2\pi_j}{\partial r_i\partial r_j}\right)\cdot\left(\frac{\partial^2\pi_i}{\partial r_i\partial r_j}\right)$$

In a stable equilibrium, this must be positive.

By applying Cramer's rule one gets

$$\frac{dr_i}{dz_i} = -\frac{\left(\frac{\partial^2\pi_i}{\partial r_i\partial z_i}\right)\cdot\left(\frac{\partial^2\pi_j}{\partial r_j^2}\right) - \left(\frac{\partial^2\pi_j}{\partial z_i\partial r_j}\right)\left(\frac{\partial^2\pi_i}{\partial z_i\partial r_j}\right)}{\left(\frac{\partial^2\pi_i}{\partial r_i^2}\right)\cdot\left(\frac{\partial^2\pi_j}{\partial r_j^2}\right) - \left(\frac{\partial^2\pi_j}{\partial r_i\partial r_j}\right)\left(\frac{\partial^2\pi_j}{\partial r_i\partial r_j}\right)} \qquad (3.1.a.**)$$

Because the following conditions are satisfied

$$\left(\frac{\partial^2 \pi_i}{\partial r_l \partial z_i}\right) = \left(\frac{\partial p_i}{\partial r_i}\right) \cdot \alpha_i \cdot p_j > 0$$

$$\left(\frac{\partial^2 \pi_j}{\partial r_j^2}\right) < 0$$

$$\left(\frac{\partial^2 \pi_j}{\partial z_i \partial r_j}\right) = -\left(\frac{\partial p_j}{\partial r_j}\right) \cdot \alpha_i \cdot p_i < 0$$

$$\left(\frac{\partial^2 \pi_i}{\partial r_i \partial r_j}\right) = -\left[(1 - z_i) \cdot \left(\frac{\partial p_i}{\partial r_i}\right) \cdot \alpha_i + \left(\frac{\partial p_i}{\partial r_i}\right) \cdot \alpha_j \cdot z_j\right] \cdot \frac{\partial p_j}{\partial r_j} < 0,$$

dr_i/dz_i is positive irrespective of the value of z_j.

Analogically $dr_i/dz_j < 0$.

The total impact of z_i on profits is

$$\frac{d\pi_i}{dz_i} = \frac{\partial \pi_i}{\partial z_i} + \left(\frac{\partial \pi_i}{\partial r_i}\right) \cdot \left(\frac{dr_i}{dz_i}\right) + \left(\frac{\partial \pi_i}{\partial r_j}\right) \cdot \left(\frac{dr_j}{dz_i}\right)$$

Because

$$\frac{\partial \pi_i}{\partial z_i} > 0, \left(\frac{\partial \pi_i}{\partial r_i}\right) = 0, \left(\frac{\partial \pi_i}{\partial r_j}\right) < 0 \text{ and } \left(\frac{dr_j}{dz_i}\right) < 0,$$

it holds that

$$\frac{d\pi_i}{dz_i} > 0.$$

Therefore, no value of z_i lower than 1 can be an optimum, and $z_i = 1$ is the dominant strategy for government i, and $z_j = 1$ is the dominant startegy for government j.

QED

Appendix 2

Result 3.1.b
In symmetric cases (ie where $\alpha_1 = \alpha_2$, $z_1 = z_2$, $\Gamma_1 = \Gamma_2$) the following holds. If both governments try to maximize social welfare, defined as the sum of domestic corporate profits and domestic consumer surplus, the only possible equilibrium entails maximal discrimination by both governments; $z_1 = z_2 = 1$.

Proof:

In the result 3.1.a it was demonstrated that if governments take into account only corporate profits, then both governments discriminate maximally. If it is possible to demonstrate that z_i (i = 1, 2) does not marginally affect domestic consumer surplus, the equilibrium for maximal discrimination will prevail even when consumer welfare is taken into account. In any symmetric case, a slightly higher value of z_i would increase the profits of domestic company i, and would not affect domestic consumers.

The expected increase in consumer utility is

$$\Gamma \cdot \left[1 - \left(1 - p_1\right) \cdot \left(1 - p_2\right)\right] \qquad (3.1.b.*)$$

where $\Gamma = \Gamma_1 = \Gamma_2$.

In a symmetric case, the marginal impact of z_i on consumer welfare in the country i equals Γ times the impact of z on the probability that the good will be available

$$\Gamma\left[\left(1 - p_j\right) \cdot \left(\frac{dp_i}{dr_i}\right) \cdot \left(\frac{dr_i}{dz_i}\right) + \left(1 - p_i\right) \cdot \left(\frac{dp_j}{dr_j}\right) \cdot \left(\frac{dr_j}{dz_i}\right)\right]. \qquad (3.1.b.**)$$

Because in a symmetric situation $p_i = p_j$, the expression (3.1.b.*) equals zero if

$$\left|\frac{dp_i}{dr_i} \cdot \frac{dr_i}{dz_i}\right| = \left|\frac{dp_j}{dr_j} \cdot \frac{dr_j}{dz_i}\right|. \qquad (3.1.b.***)$$

The equation (3.1.b.***) is satisfied if z_i affects domestic and foreign R&D equally strongly. In the light of the fact that

increasing z_i encourages domestic and discourages foreign R&D, Cramer's rule implies that this is the case if

$$\left(\frac{\partial^2\pi_i}{\partial r_i \partial z_i}\right)\cdot\left(\frac{\partial^2\pi_j}{\partial r_j^2}\right)-\left(\frac{\partial^2\pi_j}{\partial z_i \partial r_j}\right)\cdot\left(\frac{\partial^2\pi_i}{\partial r_i \partial r_j}\right)= \\ -\left(\frac{\partial^2\pi_i}{\partial r_i^2}\right)\cdot\left(\frac{\partial^2\pi_j}{\partial r_j \partial z_i}\right)+\left(\frac{\partial^2\pi_i}{\partial r_i \partial z_i}\right)\cdot\left(\frac{\partial^2\pi_j}{\partial r_i \partial r_j}\right) \qquad (3.1.b.****)$$

Because of symmetry

$$\left(\frac{\partial^2\pi_j}{\partial r_j^2}\right)=\left(\frac{\partial^2\pi_i}{\partial r_i^2}\right)$$

and

$$\left(\frac{\partial^2\pi_j}{\partial r_i \partial r_j}\right)=\left(\frac{\partial^2\pi_i}{\partial r_i \partial r_j}\right)$$

Therefore, the equation (3.1.b.****) and the whole result is valid if

$$\left(\frac{\partial^2\pi_j}{\partial r_j \partial z_i}\right)=-\left(\frac{\partial^2\pi_i}{\partial r_i \partial z_i}\right)$$

Because of symmetry, the following condition is valid

$$\left(\frac{\partial^2\pi_i}{\partial r_i \partial z_i}\right)=\left(\frac{\partial p_i}{\partial r_i}\right)\cdot\alpha_i\cdot p_j=\left(\frac{\partial p_j}{\partial r_j}\right)\cdot\alpha_j\cdot p_i=\left(\frac{\partial^2\pi_J}{\partial r_J \partial z_i}\right)$$

It follows that the value of z_i does not marginally affect expected consumer utility in country i.

QED

Appendix 3

Result 3.1.d
If either of the two countries is bigger as a market area, and if both governments try to maximize domestic R&D, an international agreement that forbids discrimination in patent policies increases R&D in the smaller country but reduces it in the big country.

Proof

We assume there is a small country 1, and a large one 2, ie $\alpha_1 < \alpha_2$. According to result 3.1.a, in the noncooperative case without agreements, the two governments will both practice maximal discrimination. Therefore $z_1 = z_2 = 1$. In the case where both governments commit themselves to nondiscrimination, $z_1 = z_2 = ½$. Therefore, this analysis can be based on the assumption that $z_1 = z_2 = z$.

The result is valid if for any value of z, $dr_1/dz < 0$ and $dr_2/dz > 0$. It will be demonstrated that this holds.

In order to simplify the analysis, the situation is re-expressed as a game where the two firms set their p-parameters instead of their R&D efforts. Because each value of p_i corresponds to one particular value of r_i, it is irrelevant to the outcome whether firms are assumed to use r_i or p_i as the decision variable. Choosing one value of r_i automatically implies choosing one value of p_i.

$$\pi_i = p_i \cdot \alpha_i \cdot \{1 - (1 - z) \cdot p_j\} + p_i \cdot \alpha_j (1 - z \cdot p_j) - C_i$$

where C_i is the cost of R&D.[12]

The optimization conditions for firm i are

$$\frac{\partial \pi_i}{\partial p_i} = \alpha_i (1 - (1 - z) p_j) + \alpha_j (1 - z p_j) - \frac{dC_i}{dp_i} = 0. \qquad (3.1.d.*)$$

The optimization conditions for firms 1 and 2 define implicitly

[12] By definition $r_i = C_i$. The notational change is due to the fact that in this analysis, R&D cost is not the decision variable but rather a function of the decision variable p_i. $C_i = C(p_i)$, $C(0) = 0$, $dC_i/dp_i > 0$, and $C(1) = \infty$.

$p_1 - p(p_2, z, \alpha_1, \alpha_2) = 0$
$p_2 - p(p_1, z, \alpha_2, \alpha_1) = 0.$

The Jacobian of this system equals

$$1 - \left(\frac{\partial p_1}{\partial p_2}\right) \cdot \left(\frac{\partial p_2}{\partial p_1}\right)$$

Applying Cramer's rule gives

$$\frac{dp_1}{dz} = \frac{-\frac{\partial p_1}{\partial z} - \left(\frac{\partial p_1}{\partial p_2}\right) \cdot \left(\frac{\partial p_2}{\partial z}\right)}{1 - \left(\frac{\partial p_1}{\partial p_2}\right) \cdot \left(\frac{\partial p_2}{\partial p_1}\right)}. \qquad (3.1.d.^{**})$$

Implicit differentiation of the optimization condition (3.1.d.*) yields

$$\frac{\partial p_1}{\partial z} = -\frac{[(\alpha_1 - \alpha_2) \cdot p_2]}{\left[-\frac{d^2 C_1}{dp_1^2}\right]}.$$

Because $\alpha_1 - \alpha_2 < 0$ and $\frac{d^2 C_1}{dp_1^2} > 0, \quad \frac{\partial p_1}{\partial z} < 0.$

Analogically

$$\frac{\partial p_2}{\partial z} = -\frac{[(\alpha_2 - \alpha_1) \cdot p_1]}{\left[-\frac{d^2 C_2}{dp_2^2}\right]} > 0.$$

Implicit differentiation of the optimization condition (3.1.d.*) yields

$$\frac{\partial p_1}{\partial p_2} = -\frac{[\alpha_1 - (\alpha_1 - \alpha_2)z]}{\left[\frac{d^2 C_1}{dp_1^2}\right]} < 0 \Rightarrow \frac{dp_1}{dz} < 0.$$

Analogically

$$\frac{dp_2}{dz} = \frac{-\frac{\partial p_2}{\partial z} - \left(\frac{\partial p_2}{\partial p_1}\right) \cdot \left(\frac{\partial p_1}{\partial z}\right)}{1 - \left(\frac{\partial p_2}{\partial p_1}\right) \cdot \left(\frac{\partial p_1}{\partial p_2}\right)} > 0$$

$\Rightarrow$ Irrespective of the value of z, a higher degree of mutual discrimination increases the R&D effort of firm 2 and decreases the R&D effort of firm 1.

QED

Appendix 4

Result 3.2.a
Let the situation be symmetric (ie $r_1 = r_2 = r$, $\alpha_1 = \alpha_2$, $f = ½$ and $z = 0$). $dr_1/dz < 0$ at least in cases where $\alpha' \approx 0$. $dr_1/dz > 0$ at least in cases where α' has the highest value it can have in a meaningful, symmetric equilibrium.

Firms' optimization conditions are

$$\frac{\partial \pi_1}{\partial r_1} = \left(\frac{\partial f}{\partial r_1}\right)\alpha_1 + \alpha_1' f - 1 = 0$$

$$\frac{\partial \pi_2}{\partial r_2} = \left(-\frac{\partial f}{\partial r_2}\right)\alpha_2 + \alpha_2' (1-f) - 1 = 0$$

By applying Cramer's rule, it is possible to calculate the total impact of z on r_1

$$\frac{dr_1}{dz} = -\frac{\left(\frac{\partial^2 \pi_1}{\partial r_1 \partial z}\right)\cdot\left(\frac{\partial^2 \pi_2}{\partial r_2^2}\right) - \left(\frac{\partial^2 \pi_1}{\partial r_1 \partial r_2}\right)\cdot\left(\frac{\partial^2 \pi_2}{\partial z \partial r_2}\right)}{\left(\frac{\partial^2 \pi_1}{\partial r_1^2}\right)\cdot\left(\frac{\partial^2 \pi_2}{\partial r_2^2}\right) - \left(\frac{\partial^2 \pi_1}{\partial r_1 \partial r_2}\right)\cdot\left(\frac{\partial^2 \pi_2}{\partial r_1 \partial r_2}\right)}$$

This is positive (negative) when the numerator is negative (positive).

When $r_1 = r_2 = r$ and $z = 0$, the numerator of this expresion is

$$N = \frac{\left[\alpha^2\{1-Ln(2)\} + r\alpha\left[\{Ln(2)-3\}\alpha' + r\{Ln(4)-2\}\alpha''\right] + r^2\alpha'\{Ln(4)\alpha' + rLn(16)\alpha''\}\right]}{\left(16r^3\right)}$$

$$\lim_{\alpha' \to 0} N = \frac{\left[\alpha^2\{1-Ln(2)\} + r^2\alpha\{Ln(4)-2\}\alpha''\right]}{\left[16r^3\right]} > 0 \Rightarrow \frac{dr_1}{dz} < 0$$

$$\Rightarrow \text{If } \alpha' \approx 0, \ \frac{dr_1}{dz} < 0.$$

In a symmetric equilibrium $z = 0 \Rightarrow \partial f/\partial r_1 = r_2/(r_1 + r_2)^2$; $r_1 = r_2 = r \Rightarrow$ $\partial f/\partial r_1 = 1/(4r)$; $f = ½$.

$\Rightarrow$ $\partial \pi_i/\partial r_i = \alpha/(4r) + \alpha'/2 - 1 = 0 \Rightarrow r = \alpha/(4-2\alpha')$, implying that in no meaningful symmetric equilibrium is $\alpha' \geq 2 \Rightarrow$ The highest possible value of α is 1.999999 ... (If $\alpha'' << 0 \Rightarrow$ the second order condition holds.)

By substituting $\alpha/(4-2\alpha')$ for r in the expression for N, it can be calculated that:

$$\lim_{\alpha' \to 2} N = \frac{\mathrm{Ln}(16)\alpha''}{8} < 0 \Rightarrow \frac{dr_1}{dz} > 0$$

$\Rightarrow$ If $\alpha' \approx 2, \frac{dr_1}{dz} > 0.$

QED

4 Use of patents as costly insurance: A model to explain empirical observations

4.1 Introduction

Data availability problems have probably affected greatly the quantity and composition of empirical studies on R&D. Access to firm-level data concerning R&D expenditures or research personnel is often very limited, whereas patent documents are public. In Finland, satisfactory patent data are available for the period since 1840, but there are no systematically collected statistics on R&D expenditures or research personnel for the pre-1969 period.

But are patent counts a satisfactory proxy for technological progress? As they have been used over and over again for this purpose, the relevance of dozens of empirical papers hinges entirely on the correct answer to this question.

In the following, a simple model on the patenting behaviour of rival duopolists will be presented. A firm finds it reasonable to patent its invention if it is likely that a rival might invent the same technology; patenting is needed to safeguard the monopoly profits that would otherwise be jeopardized by the rival firm and its R&D laboratories. If no such competitive pressure exists, the invention can be monopolized more effectively through secrecy.

The model provides two testable hypotheses:

1) Patent counts are a satisfactory proxy for R&D efforts, except in the short run at the firm level.
2) Firms are more willing to patent their inventions during recessions than during booms.

Both of these hypotheses receive support from empirical observations.

4.2 Previous literature

4.2.1 Theoretical contributions

In the previous literature, both the societal (dis)advantages of the patent system and the properties of optimal patent legislation have been a central topic for numerous theorists (See Reingaum 1982, Fudenberg *et al* 1983, Muto 1987, Lippman and McCardle 1988, David and Olsen 1992, Hausman and MacKieMason 1992, de Fraja 1993, Aoki and Prusa 1993 and Horowitz and Lai 1996, to mention a few). In these papers, the general assumption has been that it is always optimal to patent.

In addition, there are some models that are used to analyse the "to patent or not to patent" dilemma. In these papers, two problems related to patents are taken into account:

1) Patents do not guarantee perfect monopoly rights
2) Patents reveal information.

Horstmann *et al* (1985) presented a duopoly model on patents as an information transfer mechanism; by patenting, the innovator reveals information about the likely profitability of the invention. Paradoxically, firms may be more willing to patent their inventions if future new technologies based on the invention would be of limited value.

In the model of Choi (1990), a firm guarantees itself monopoly rights by patenting its invention, but it also reveals valuable technological information that may be of use to other firms in developing a more advanced technology. Firms are unwilling to patent if there are several competitors and if the eventual inventions based on the new technology are likely to be valuable.

In Saarenheimo's (1994) model, the innovator must decide whether to patent or to keep the invention secret. Patenting enables the firm to license its innovative output, but the resulting knowledge spillovers may help other firms to reach a more advanced stage of technology.

Takalo (1996) analyses how patent length and breadth affect firms' incentives to patent their inventions or to keep them secret. The main focus is on the impact of legislation on the incentives to patent. Patent breadth increases the incentives to patent whereas patent life may have the opposite effect. An optimal patent policy should prefer short patents.

4.2.2 Empirical observations yet to be explained

Now, we shall briefly present three stylized facts concerning the correlation between patent counts and R&D efforts. The following stylized facts provide the empirically related motivation for the model to be presented in the following.

Stylized fact 1: There is a positive short-term correlation between patenting and R&D at both the industry level and the macroeconomic level. When an industry or a nation increases its R&D investments, the number of patent applications filed increases almost immediately, or at least with a relatively short lag.

Using panel data, Pakes and Griliches (1984) found that the number of patent applications changed over time in accord with technology efforts at the industry level. There did not even seem to be any significant lag between an increase in R&D efforts and the resulting increase in the number of patent applications filed.

Panel estimations by Hall *et al* (1986) revealed a clear statistical interrelationship between R&D expenditure and patent applications filed at the industry level. The effect seemed to be immediate: there was no lag between change in R&D effort and change in number of applications filed.

Trajtenberg (1990) has analysed the history of the computer tomography industry. Using longitudinal data describing the whole industry, he found that the total sum of money spent on R&D does not explain adequately the economic importance of patented innovations, but it explains the total number of patents much better.

Griliches (1989, reviewed in Griliches 1990) analysed the determinants of domestic patent applications using longitudinal macro-level data from US. Aggregate real R&D expenditures increased the number of applications filed in the following year, albeit less than proportionately.

Stylized fact 2: In the long term, there is a clear positive correlation between R&D effort and the number of patents applications, irrespective of whether the data are from the firm, industry or country level.

Using US data, Bound *et al* (1984, p. 41) found a statistically significant relationship between patents filed in 1969–1979 and R&D expenditures at the firm level. R^2 for the OLS estimations varied between 0.65 and 0.77, when controlled for industry-specific effects. The explained variable was not adjusted for firm size.

In the US pharmaceutical industry, there is a substantial firm-level correlation (0.6–0.8) between patents filed over the years 1975–1982

and several other measures of corporate excellence in science and technology, including subjective expert evaluations and bibliometric data (Narin *et al* 1987, p. 151).

Using a cross-sectional sample, Schmookler (1966, p. 44) found that about 85 % of inter-industry variation in US patenting could be explained by variation in R&D expenditure. Using similar but newer statistics, Pavitt (1982) found that the propensity to patent was disproportionately low in certain industries. But when these industries were excluded, the correlation between R&D expenditure and patents was clear, varying between 0.71 and 0.82. The results of Crépon and Duguet (1997) corroborate this view, based on data covering 181 French firms.

Stylized fact 3: The statistical connection between R&D and patent counts is weaker at the firm level than at the industry level, probably at least partly because the number of patent applications filed by a company depends not only on its own R&D effort but also on that of its rivals. This phenomenon is especially clear if one considers short-term variations in the two variables.

According to Pakes and Griliches (1984), the degree of temporal covariation between R&D and patents is lower at the firm level than at the industry level. Only about 25 % of within firm variance in patenting could be explained by changes in the R&D expenditures of the respective firms (p. 65). The authors suggested that this might simply reflect the relatively higher degree of randomness in a data set collected at a more disaggregated level, although they did not rule out the possibility of other explanations.

In India, there is no immediate interrelationship between patents and R&D effort at the firm level, at least not when controlled for number of employees with a PhD (Deolalikar and Röller, 1989).

In the log-linear OLS results of Acs and Audretsch (1989), the R&D effort of the firm explained the number of patent applications filed by the same firm only slightly better than did the R&D effort of the whole industry. The data set was a cross-sectional sample of US firms covering a period of one year. These results may reflect spillover effects between firms, but nothing rules out other explanations.

Cincera (1997) applied four different count model estimation techniques to analyse the impact of own and rival R&D on the number of patent applications filed at the company level. The estimations were carried out using a panel data set covering 181 companies from different parts of the world from various industries. The results seem to depend relatively strongly on the estimation method. Nevertheless, at least the following findings are interesting. In terms of the estimated

coefficient, the *immediate* impact (with no lag) of rival R&D on the number of patent applications filed by a firm was stronger than the impact of R&D performed by the company itself, irrespective of the estimation method. As to the degree of statistical significance, rival R&D had a higher degree of statistical significance in three cases out of four. Secondly, in two specifications out of four, rival R&D effort had a stronger total impact than own R&D effort on the number of patent applications, when all lagged values of explanatory variables were taken into account.

According to Devinney (1993), patent statistics measure innovative output relatively well at the industry level (and better than at the firm level). The data covered a period of 14 years. According to the estimations, the number of new products has a statistically significant relation to patenting intensity, even at the firm level, but this significance is mainly due to the huge size of the sample (3033 US firms), rather than to a good regression fit (R^2 = 0.02). Surprisingly, the number of new products introduced by individual firms covariated stronger with patenting by the whole industry than with patenting by the firm itself. It is difficult to explain this observation with arguments based on knowledge spillovers between firms.

Stylized facts 1 and 2 may not seem especially surprising, but certainly fact 3 does. How could it be possible that when a firm increases its R&D effort, and possibly even its R&D output, it does not file more patent applications, but when rivals increase their R&D efforts the firm increases its filings. Even in the long run, as demonstrated by Devinney (1993), the number of patent applications filed by a firm reflects the R&D output of the whole industry, not only the innovative activities of the firm itself. The following model is, above all, an attempt to explain this observation.

4.3 The model

4.3.1 Assumptions

This model describes a technology-intensive duopoly where only monopolized inventions enable firms to earn profits. Inventions can be monopolized either with patents or by keeping them secret. A basic empirical fact behind the following argumentation is that not all inventions are patented (Mansfield 1986, p. 177).

There are two firms in this model (1 and 2) with an exogenously given capability to innovate. They both try to develop the same product. In this attempt, they may either succeed or fail. Both firms can freely choose any probability (p_i) of succeeding in the product development effort. However, increasing the probability requires costly R&D. The cost, denoted r_i, is most readily interpretated as R&D expenditure.

Firms compete in the following game:

Firms observe the values of two random exogenous factors, t_1 and t_2. The factor t_i directly affects firm i's cost of achieving any given success (p_i). This parameter could be thought of as government subsidies granted more or less arbitrarily, good or bad luck in R&D, qualifications of R&D work force or unevenly distributed knowledge spillovers from, say, basic academic research. The value of the exogenous parameter is not necessarily the same for both firms.

Firms choose the probabilities of success in product development (p_1, p_2). Firms know everything that affects the rival's optimal choice and they are able to calculate each other's probability of success. The cost (r_i) of achieving a given probability of success (p_i) is a function only of the probability itself and the exogenous parameter t_i. Any value of p_i that satisfies $0 \leq p_i < 1$ can be chosen, provided the firm is willing to pay the R&D cost (r_i) that corresponds to the chosen level of p_i.

The following conditions describe the cost as a function of the probability to innovate.

– $p_i = 0$ can be achieved at no cost; $p_i = 0$ implies $r_i = 0$.

– Increasing the probability increases the cost, and the marginal cost is increasing;

$$\frac{\partial r_i}{\partial p_i} > 0; \quad \frac{\partial^2 r_i}{\partial p_i^2} > 0.$$

– It is not possible to guarantee success at any finite cost

$$\lim_{p_i \to 1} r_i = \infty$$

Both firms observe whether own R&D investment is successful or not, but they cannot observe whether the rival's R&D effort is successful. Instead, they are both aware of the probability that the rival has

innovated (p_i). The success or failure of a firm tells nothing about the success or failure of its rival; p_1 and p_2 are independent probabilities.

If either (or both) of the two firms has been successful and has made the invention, the innovator(s) decide(s) whether to apply for a patent. The alternative of immediate patenting has two different interpretations:

1) The patent application will never be filed.
2) Filing the patent application will be postponed.

We shall return to these two potential interpretations of the model later. The alternative of immediate patenting is now simply called non-patenting.

When making its patenting decision, the firm still does not know whether the rival has made the same invention or not. Instead, both firms are completely aware of the factors that affect the incentives of the rival to file a patent application, provided there is something to be patented.

If at least one patent application is filed, public authorities choose the patentee. If there is only one application, the applicant automatically becomes the patentee. If two applications are filed, the patentee is chosen at random, and both firms have the same probability of getting the patent.

The invention is commercialized.

Due to the exogenous factors t_1 and t_2 affecting R&D incentives, *any combination of p_1 and p_2 would be implemented by some values of these exogenous factors*.

This model aims at explaining firms' patenting policies, not R&D investments. In the following, the probability of success parameters (p_i) are mainly treated as being exogenous, the focus being on the patenting behaviour implemented by different quasi exogenous combinations of p_1 and p_2. However, we shall also briefly discuss the first stage of the game in a separate section.

Only a monopolized invention can enable firms to earn revenue. Thus, unless the firm succeeds in developing the invention, its profits at the commercialization stage are zero. Not even an innovator will earn a return if it does not manage to monopolize the invention. This can happen in three different cases:

1) If both firms make the invention but neither tries to patent it, the firms end up in Bertrand competition, and neither can earn positive revenue on the invention.

2) The inventor gets no profits if it does not file a patent application but the rival does file. The rival will acquire a legal monopoly.
3) If both firms file a patent application, the loser of the patent contest earns nothing for the invention. The rival acquires a legal monopoly.

An invention can be monopolized more or less successfully. The commercial value of the invention is either α or β, depending on the patenting strategy and the rival's actions.

If the firm manages to get a patent, the value of the invention equals α. It does not matter whether the firm gets the patent because it is the only applicant or because it wins the patent contest. In either case, the value of the patented invention is α. $\alpha > 0$.

The earnings level β is reached if the innovator does not try to patent the invention and the rival has not made the invention. The monopoly position of the firm is based on keeping the essential details of the invention secret.

One quite natural interpretation of the β-parameter is that it is the full monopoly return on the invention. In real life, it has often been rather easy to invent around patents. Such imitations affect negatively the profits of the patentee. Patent applications are public documents. Because they are an excellent source of information for imitators, the monopoly profits are higher if the invention is monopolized through secrecy instead of patenting. Here, it is assumed that there is a large number of non-innovative firms that can create imperfect substitutes for the invention if the inventor publishes crucial details concerning the invention by filing a patent application. Imitating is assumed to be a perfectly competitive industry, and because of free entry, no profits can be earned in such activities. The two firms described in this model may or may not imitate each other's inventions, but even if they do, these activities have no impact on their payoff functions. Imperfect imitations will cause the patentee significant losses, and therefore $\beta > \alpha$.

The assumption that inventions monopolized through secrecy are more valuable than inventions monopolized via patents is consistent with empirical findings. In the sample of Mansfield *et al* (1981), 60 % of patented innovations were imitated within 4 years. According to Mansfield (1986), only in the pharmaceutical and chemical industries, did interviewed managers regard patents as essential to the development or introduction of inventions in more than 30 % of cases. According to Levin *et al* (1987), interviewed managers estimated that secrecy was slightly more effective than patenting as a means to keep

process innovations secret, whereas in the case of product inventions, patenting was regarded as slightly more effective than secrecy.

4.3.2 Solving the model

4.3.2.1 How do firms choose their patenting strategies

The patenting stage can be described with a classical game matrix where there are two players, both having two possible strategies: to patent or not to patent. The situation differs from the standard textbook game model in the sense that there is a certain probability that a player does not participate: if a firm has not invented, it certainly cannot patent, and the question of its optimal patenting strategy is irrelevant.

However, a firm cannot know whether its rival has succeeded or not. When it makes its own patenting decision, it knows nothing but the probability that the rival has invented. It is also able to calculate the optimal patenting decision of the rival. The situation is completely analogous with a game where the firms would have to commit themselves to a given patenting strategy *before* they can observe the results of their own R&D efforts. Hence, we can assume that the two firms have to choose a patenting strategy in any case, even though in many cases, the decision is completely irrelevant.

The strategy where the firm files a patent application if it has made the invention is denoted P, and the non-patenting strategy is denoted non-P.

If firm i has invented and if both firms apply the P-strategy, firm i will find itself in either of two situations. The first situation is that the rival (j) has invented and that firm i has a 50 % chance to win the patent contest and to earn a profit of α. The second situation is that the rival has not invented and that firm i is the sole patent applicant and will earn a profit of α. The probability that the rival has invented is p_j. Hence the expected payoff is $p_i \cdot \alpha[½p_j + (1-p_j)]$.

If the rival firm (j) applies the non-P strategy and firm i applies the P-strategy, firm i will certainly earn a profit of α, provided it has invented. Hence the expected payoff is $p_i \cdot \alpha$, and the probability that the rival has invented (p_j) can be ignored.

If firm i decides to apply the non-P strategy, it will earn a profit (β) if and only if firm j has not invented. The expected payoff is $p_i \cdot (1-p_j) \cdot \beta$. In this case, the rival's patenting strategy is irrelevant.

When firm i chooses its patenting strategy, the probability that it innovates itself (p_i) has no direct impact on the relative profitability of the two alternative strategies at the patenting stage, at least not if the rival's patenting strategy is regarded as exogenous. Hence, when analysing the patenting decision to be made by firm i, the variable p_i is of no relevance. Therefore, we can express the payoffs without the probabilities that the firm itself innovates. The traditional game matrix for the situation is as follows:

	Firm 1: P	Firm 1: Non-P
Firm 2: P	$½\cdot p_2\cdot\alpha+(1-p_2)\cdot\alpha$ $½p_1\cdot\alpha+(1-p_1)\cdot\alpha$	$(1-p_2)\cdot\beta$ α
Firm 2: Non-P	α $(1-p_1)\cdot\beta$	$(1-p_2)\cdot\beta$ $(1-p_1)\cdot\beta$

There are four possible subgame perfect equilibria:

1) {P,P}. This equilibrium cannot prevail unless

$$½\cdot p_2\cdot\alpha + (1-p_2)\cdot\alpha > (1-p_2)\cdot\beta \Leftrightarrow p_2 > (2\beta-2\alpha)/(2\beta-\alpha)$$

and

$$½p_1\cdot\alpha + (1-p_1)\cdot\alpha > (1-p_1)\cdot\beta \Leftrightarrow p_1 > (2\beta-2\alpha)/(2\beta-\alpha).$$

If this equilibrium prevails, P is a dominant strategy for both firms.[13]

2) {P,non-P}. This equilibrium cannot prevail unless

$$\alpha > (1-p_2)\cdot\beta \Leftrightarrow p_2 > (\beta-\alpha)/\beta$$

and

$$½p_1\cdot\alpha + (1-p_1)\cdot\alpha < (1-p_1)\cdot\beta \Leftrightarrow p_1 < (2\beta-2\alpha)/(2\beta-\alpha).$$

[13] Proof:
The equilibrium P–P prevails when the following condition is satisfied, whether i = 1 or i = 2: $½pi\cdot\alpha + (1-pi)\cdot\alpha > (1-pi)\cdot\beta \Leftrightarrow (1-½pi)\cdot\alpha > (1-pi)\beta$ which implies $\alpha > (1-pi)\beta \Rightarrow$ firm j chooses P even if firm i choses non-P.

3) {non-P,P}. This equilibrium cannot prevail unless both

$\alpha > (1-p_1)\cdot\beta \Leftrightarrow p_1 > (\beta-\alpha)/\beta$

and

$½p_2\cdot\alpha + (1-p_2)\cdot\alpha < (1-p_2)\cdot\beta \Leftrightarrow p_2 < (2\beta-2\alpha)/(2\beta-\alpha)$.

4) {non-P,non-P}. This equilibrium cannot prevail unless

$\alpha < (1-p_1)\cdot\beta$ and $\alpha < (1-p_2)\cdot\beta \Leftrightarrow p_2 < (\beta-\alpha)/\beta$; $p_1 < (\beta-\alpha)/\beta$.

Any of these four possible outcomes could prevail, provided the variables α, β, p_1 and p_2 have suitable values.

It is certain that there is always at least one subgame perfect equilibrium.

Result 4.3.a
It is not possible to construct examples where there would be no subgame perfect equilibrium at the patenting stage.

Proof:

Appendix 1.

The outcome as a function of the innovation probabilities can be described as in figure 11. Each point in the square corresponds to a certain combination of p_1 and p_2, and the resulting combination of patenting strategies is given in parentheses, the strategy of firm 1 being stated first.

Figure 11. **Patenting strategies with different innovation probabilities**

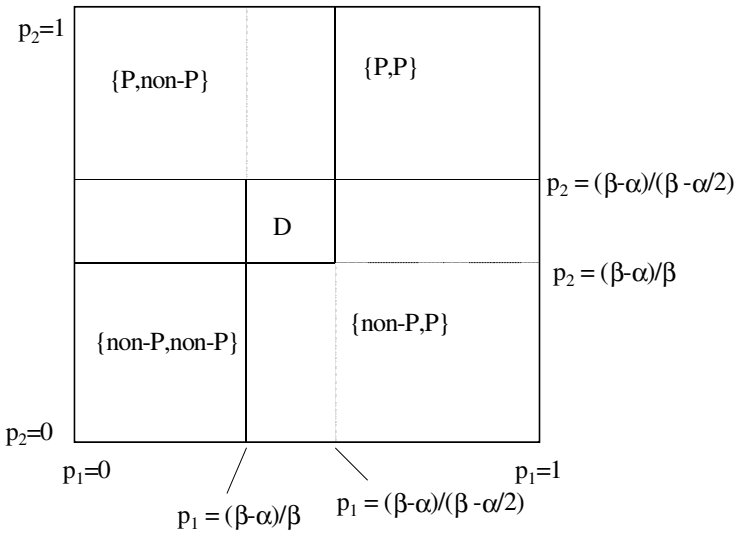


Figure 11 The patenting stage outcome as a function of p_1 and p_2. In the D-area, there are two possible subgame perfect equilibria.

Another interesting question is the uniqueness of the equilibrium. Some types of multiple equilibria can be ruled out.

Result 4.3.b
If either {P by both firms} or {non-P by both firms} is a subgame perfect equilibrium, there cannot be any other subgame perfect equilibria.

Proof:

Appendix 2.

However, it is possible to find other cases where there are two subgame perfect equilibria. Double equilibria situations where either firm 1 plays P and firm 2 non-P, or eventually *vice versa*, cannot be ruled out. Such double equilibria prevail in the area denoted D in figure 11. In these cases, the following two conditions must be satisfied simultaneously for both p_1 and p_2:

1) If the rival i plays non-P, the strategy P by firm j must yield a higher expected profit than the strategy non-P. Therefore $\alpha > (1-p_i)\cdot\beta$.
2) If the rival i plays P, the strategy non-P by firm j must yield a higher expected profit to firm j than the strategy non-P. Therefore $(1-p_i/2)\alpha < (1-p_i)\cdot\beta$.

These conditions are satisfied if $(\beta-\alpha)/(\beta-\alpha/2) > p_1 > (\beta-\alpha)/\beta$ and $(\beta-\alpha)/(\beta-\alpha/2) > p_2 > (\beta-\alpha)/\beta$.

The existence of multiple equilibria is analytically somewhat problematic. Fortunately, cases where there would be two subgame perfect equilibria are not very likely. According to simple calculations, the difference between $(\beta-\alpha)/\beta$ and $(\beta-\alpha)/(\beta-\alpha/2)$ is maximized when α equals $(2-\sqrt{2})\cdot\beta \approx 0.586\cdot\beta$. Even in this extreme case, we can easily calculate that only about 3 % of possible combinations of p_1 and p_2 would imply a double equilibria situation. With any other ratio of α to β, the share of such cases would be smaller. Hence, the analytically problematic double equilibria outcome is highly unlikely.

In fact, one might argue that the probability that the firms would end up in the D-area is even lower, if firms behave strategically at the R&D stage. In double equilibria cases, the expected profit of player i, who chooses strategy P, is $p_i\alpha$, and the expected profit of player j, who plays non-P, is $p_j\beta(1-p_i)$. Because two equilibria cannot co-exist unless $\alpha > (1-p_i)\cdot\beta$, the P-player has a higher expected profit than the non-P player, at least if $p_i = p_j$. Therefore, both firms would prefer to be the player who chooses P. Nevertheless, if the rival plays P, it is better to play non-P.[14] If either of the firms knows at the R&D stage that it will be the non-P player, and thus probably earn a lower profit than the rival, the firm might artificially restrict its likelihood to innovate. This would be a strategic commitment to be the P-player. If firm 1 applied this strategy, it would move the combination of p's leftward from the D area to the {P, non-P} area in figure 11. Even though this would restrict the likelihood that the firm would earn a profit, it would in many cases increase the expected value of the profit. This could be true even if one assumed that restricting the

[14] If firms are allowed to apply mixed strategies, there is a third subgame perfect equilibrium as well. In this third equilibrium, both firms play mixed strategies and both have the same expected payoff as the non-P player in a pure strategies equilibrium.

likelihood to innovate would not enable the firm to economize in its R&D budget (proof not shown here).

Because there is always at least one equilibrium, and the existence of multiple equilibria is unlikely, the two firms would normally end up in a situation where there is only one subgame perfect equilibrium at the patenting stage.

4.3.2.2 Number of patents as a function of innovation effort

This model was developed to explain certain empirical regularities in the correlation between R&D effort and patent count. Even though the number of patent applications in the model is either 1 or 0, the model does generate empirical predictions concerning aggregate data.

The direct contribution of a firm to the aggregate number of patent applications in an industry depends on two factors: the number of its innovations and its propensity to patent. The expected value of the aggregate number of patent applications is the sum of expected values of the number of patent applications for each firm included in the analysis. If the sample consists of a very large number of firms, the observed number of patent applications is not likely to differ substantially from its expected value. Thus, from the point of view of empirical predictions, it is of interest to study how the expected number of patent applications depends on the probability of innovating (p_i).

Now let us consider how the values of p_1 and p_2 affect the expected value of the number of patent applications filed by the two firms. Interestingly, it can be demonstrated that patenting by a firm may reflect the innovative effort of its rival rather than its own effort.

Result 4.3.c
Let X and Y each be a pair of probabilities p_i and p_j so that $p_i^x = p_i^y$ but $p_j^x > p_j^y$. Let the value of p_i be high or low enough to make firm j's decision independent of firm i's patenting strategy. Firm i may then have an incentive to change from Non-P to P but not from P to Non-P if p_j increases from p_j^x to p_j^y.

Proof:

Subresult 4.3.c.1
If firm j applies the P strategy and firm i is indifferent between patenting and non-patenting, then a higher value of p_j would

induce the firm i to adapt the P strategy and a lower value of p_i to adapt the non-P strategy.

Proof:

Because firm i is almost indifferent between patenting and non-patenting, $(1–½p_j)\alpha \approx (1–p_j)\cdot\beta$.

With higher values of p_j, this equation would no longer hold.

Differentiating the left hand side of this equation with respect to p_j yields $–½\alpha$. The derivative of the right hand side is $–\beta$.

On the other hand, $–\beta < –\alpha/2$. Therefore, for a higher value of p_j, $(1–p_j/2)\cdot\alpha > (1–p_j)\beta$.

Therefore, more rival R&D induces firm i to prefer the P strategy, ie patenting decisions are strategic substitutes. Analogically, less rival R&D induces firm i to prefer the non-P strategy.

QED

Subresult 4.3.c.2
If firm j applies the non-P strategy and firm i is indifferent between patenting and non-patenting, a higher p_j would induce firm i to adapt strategy P and a lower value of p_j to adapt the non-P strategy.

Proof:

Indifference implies $\alpha \approx (1–p_j)\beta$. Again, differentiating α with respect to p_j yields 0. Differentiating $(1–p_j)\beta$ with respect to p_j yields $–\beta$, which is negative. It follows that higher rival R&D induces firm i to prefer the P strategy.

QED

Hence intense rival R&D may be reflected in a strong propensity to patent. The intuition is fairly simple. When the firm does not face much technological competition, no one is likely to rediscover independently the same technology. Hence, the legal monopoly obtained by patenting is not needed. The innovator prefers to deter imitations by keeping the innovation secret and to pursue maximal profits (β).

On the other hand it is not obvious that increasing the firm's own R&D will always lead to a higher expected value of the number of patent applications filed. If the firm patents in any case, the relationship between inventing and patenting is as simple as it is often assumed to be. If the rival (j) has very low R&D expenditure, firm i will not patent anything, and the expected value of the number of patent applications will equal zero.

In some cases, a greater invention effort may make the firm unwilling to patent. This curious result is caused by rival reactions. If firm j prefers non-patenting with a given firm i R&D effort, a higher R&D effort by firm i could make firm j prefer patenting. This, in turn, could make firm i prefer non-patenting.

Result 4.3.d
If $[2\cdot\beta-2\cdot\alpha]/[2\cdot\beta-\alpha] > p_j > [\beta-\alpha]/\beta$, the following holds. The expected value of the number of patent applications filed by firm i is zero if the probability that it innovates is close to +1. But if the probability is close to zero, the expected value of the number of patent applications filed by the firm is positive.

Proof:

If $p_i < (\beta-\alpha)/(\beta)$, firm i plays P. The expected value of the number of patent applications filed by firm i equals $p_i > 0$.

If $p_i > (2\beta-2\alpha)/(2\beta-\alpha)$, the expected number of patent applications is zero, because firm i plays non-P (see figure 11).

QED

However, one should remember that this extreme result is valid if and only if the innovative effort of the rival firm (j) is within a certain relatively narrow band, when $[2\cdot\beta-2\cdot\alpha]/[2\cdot\beta-\alpha] > p_j > [\beta-\alpha]/\beta$.

4.3.3 Countercyclical patenting by decreasingly risk-averse firms

The assumptions of the basic model are now revised as follows:

1) Firms operate for two periods.

– During the first period, called the R&D period, firms engage in R&D.
– During the second period, called the commercialization period, firms may patent, and they commercialize the research results.

2) In addition to the profit earned with an invention, the firms have an exogenous flow of revenue. The value of the exogenous revenue (x_i) is observed and earned during the commercialization period but before firms decide their patenting policies. Each firm can observe the other's x value and hence takes it into account deciding on patenting. They cannot take each other's x into account when they decide on R&D efforts. As to real life interpretations, the parameter x could denote either the profit from a project of the firm itself, or eventually from other sources of revenue.

3) The total utility of firm i is the sum of two sub-utilities, one for the R&D period (W), and another one for the commercialization period (U):

$$W_i[r_i] + d \cdot U_i(\pi_i + x_i)$$

where d = the discount factor, π_i = the revenue earned with the innovation (0, α or β), x_i = exogenous revenue of the commercialization period, and r_i = R&D expenditure ($dW_i/dr_i < 0$). The subutility function U has the property of decreasing absolute risk aversion.

In the following, we focus on the impact of the x parameters on firms' patenting decisions. When the firms decide whether to patent, the utility for the R&D period has already been determined and can no longer be affected. Because the two sub-utility functions are separated, the value of W does not affect the risk aversion of the commercialization period. Hence only the variables that enter the sub-utility function, U, are of interest for the patenting decision.

With these modified assumptions, the game matrix appears as shown below. Basically the payoff functions are affine transformations of the actual payoffs, which in turn do not differ much from those in the basic model characterized by risk neutrality.

	Firm 1:P	Firm 1: non-P
Firm 2: P	$(1-½\cdot p_2)\cdot U\{\alpha+x_1\}+½p_2\cdot U\{x_1\}$ $(1-½\cdot p_1)\cdot U\{\alpha+x_2\}+½p_1\cdot U\{x_2\}$	$(1-p_2)\cdot U(\beta+x_1)+p_2U(x_1)$ $U\{\alpha+x_2\}$
Firm 2: Non-P	$U\{\alpha+x_1\}$ $(1-p_1)\cdot U(\beta+x_2)+p_1U(x_2)$	$(1-p_2)\cdot U(\beta+x_1)+p_2U(x_1)$ $(1-p_1)\cdot U(\beta+x_2)+p_1U(x_2)$

Interestingly, decreasing absolute risk aversion has a strong tendency to induce the firm to prefer patenting to non-patenting, when the firm is worse off.

Result 4.3.e
Assume that firm i is indifferent between strategies P and non-P. The rival patenting decision is exogenously determined. A higher value of the exogenous revenue, x_i, induces firm i to prefer the non-P strategy. A lower value of x_i would induce it to prefer the P-strategy.

Proof:

First, it will be proved that the result is valid if the rival (j) applies strategy P.

Because firm i is indifferent between the two strategies,

$$\left(1-½p_j\right)\cdot U\{\alpha+x_i\}+½p_j\cdot U\{x_i\}\approx\left(1-p_j\right)\cdot U\{\beta+x_i\}+p_j\cdot U\{x_i\}.$$

Rearranging terms yields

$$U\{\alpha+x_i\}=\left[\frac{\left(2-2p_j\right)}{\left(2-p_j\right)}\right]U\{\beta+x_i\}+\left[\frac{p_j}{\left(2-p_j\right)}\right]\cdot U\{x_i\}.$$

This equation could also describe a situation where firm i has risk-free wealth worth $x_i + \alpha$ and is indifferent between participating and not participating in a lottery characterized as follows:

- The firm has a $[(2–2p_j)/(2–p_j)]$ chance to win and a $[p_j/(2–p_j)]$ chance not to win.
- The price of a lottery ticket is α.
- Initial wealth is $x_i + \alpha$.
- The prize is β.
- The firm is indifferent between participating and not participating.

Decreasing absolute risk aversion implies that an exogenous increase in the exogenous wealth parameter, x_i, induces the to firm prefer participating in the lottery. With a slightly higher value of x_i,

$$U\{\alpha + x_i\} < \left[\frac{(2 - 2p_j)}{(2 - p_j)}\right] U\{\beta + x_i\} + \left[\frac{p_j}{(2 - p_j)}\right] \cdot U\{x_i\}$$

and

$$(1 - \tfrac{1}{2}p_j) \cdot U\{\alpha + x_i\} + \tfrac{1}{2}p_j \cdot U\{x_i\} < (1 - p_j) \cdot U\{\beta + x_i\} + p_j \cdot U\{x_i\},$$

so that the firm would prefer the non-P strategy.

If, instead, the rival applies the strategy non-P, the proof is even simpler. If firm i is indifferent between P and non-P, then

$$U\{\alpha + x_i\} = (1 - p_j) \cdot U(\beta + x_i) + p_j U(x_i).$$

This equation could also describe a lottery characterized as follows:

The firm has a risk free wealth of $\alpha + x_i$.
The prize is β.
The probability of winning is $(1-p_j)$.
The price of a lottery ticket is α.
The firm is indifferent between participating and not participating.

Decreasing risk aversion implies that an increase in the exogenous parameter x induces the firm to participate and a decrease induces it not to participate.

QED

This result is quite intuitive. In the model, patenting is a form of insurance that protects the firm against imitation. If the firm becomes less risk averse, it obviously begins to lean toward non-patenting.

Another interesting question is the eventual impact of the rival's exogenous revenue on the patenting propensity. As in the previous

version, firms' patenting decisions can under certain conditions affect each other's choices. Thus the exogenous revenue, x, of a firm can affect the patenting behaviour of its rival. As can be seen in the following, a firm might indirectly react to high rival profits by non-patenting.

Result 4.3.f
Let X_h and X_l be two possible values of x_j, $X_h > X_l$. It is possible that firm i chooses the P strategy if x_j has the higher value X_h, and the non-P strategy if x_j has the lower value X_l.

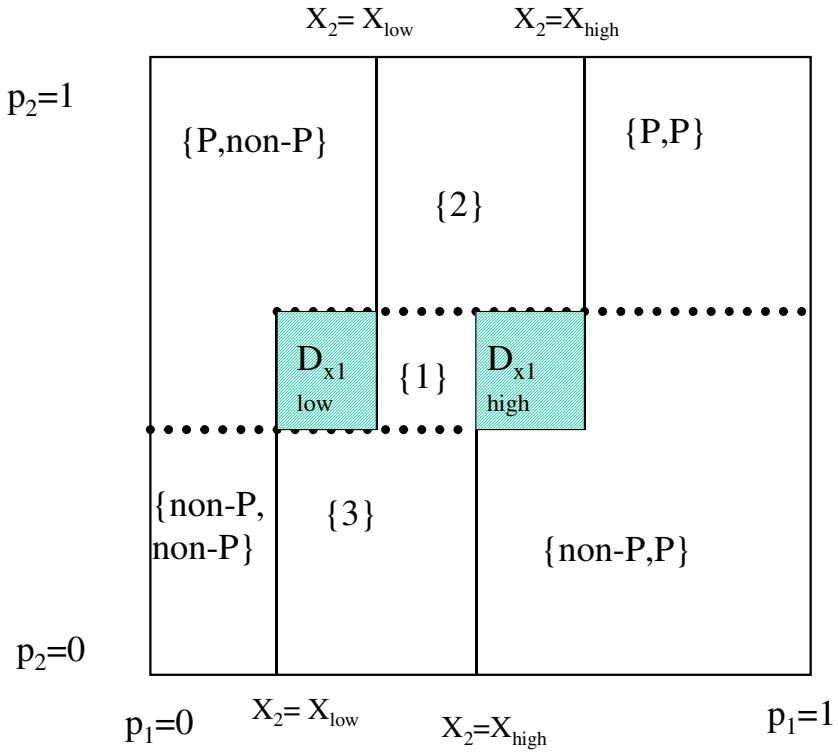
Proof:

Compare two examples of this game. In the first $x_j = X_l$, and in the second $x_j = X_h$. Otherwise the two games are similar. In the first game, firm j prefers the P strategy while firm i prefers non-P.

The value of x_j is X_h in the second game, and firm j becomes less risk averse and shifts to the strategy non-P. This does not affect the expected utility provided by the non-P strategy for firm i, but it increases the expected utility provided by the P-strategy for firm j. Firm i may therefore shift to the P strategy, which induces firm j to prefer the non-P strategy even more.

QED

Figure 12. **Graphic illustration of result 4.3.f**



Result 4.3.f refers to cases in area 1. If the value of x_2 is high, firm 1 plays P and firm 2 non-P. If the value of x_2 is low, firm 1 plays non-P and firm 2 P. (With a high value of x_2 area 2 belongs to area {P, non-P} and with a low value of x_2 it belongs to area {P, P}. With a high value of x_2 area 3 belongs to area {non-P, non-P} and with a low value of x_2 it belongs to area {non-P, P}.)

Thus, if a decreasingly risk averse firm is hit by a recession (low x), it becomes more willing to patent its own research results, because it would become less willing to gamble with non-patented inventions. If its rival (j) is hit by a recession, there is no direct impact on the preferences of firm i. However, because losses make the decreasingly risk averse rival (j) even more risk averse, the rival will patent more, and firm i, not being hurt by the recession, might react to this rival policy decision by not applying for a patent for its invention.

4.3.4 The R&D stage

The aim of this model is to analyse the patenting decision, with the factors that induce a firm to prefer a given amount of R&D effort assumed to be almost exogenous. However, a few matters should be noted concerning the reaction functions in the R&D stage and the likely outcome of the R&D stage in general.

Let Ω denote the objective function of firm i. In the risk neutrality case, Ω is the expected value of the profit. The first order condition for firm i is $\partial\Omega/\partial p_i = 0$. The slope of the reaction function at the R&D stage can be calculated by implicit derivation:

$$\frac{dp_i}{dp_j} = -\frac{\left[\dfrac{\partial^2\Omega}{\partial p_i \partial p_j}\right]}{\left[\dfrac{\partial^2\Omega}{\partial p_i^2}\right]}. \qquad (4.3.i)$$

Because in the case of a risk neutral firm the denominator is always the same,

$$\left(\frac{\partial^2\Omega}{\partial p_i^2}\right) = -\frac{\partial^2 r_i}{\partial p_i^2}.$$

This is negative by assumption, and therefore the whole expression 4.3.i is positive if $\partial^2\Omega/\partial p_i\partial p_j > 0$, and negative if $\partial^2\Omega/\partial p_i\partial p_j < 0$. As to different combinations of patenting strategies, the following four possibilities obtains.

Case 1: If both firms follow the P-strategy, then with risk neutrality $\Omega = [½p_j\alpha + (1-p_j)\cdot\alpha]\cdot p_i - r_i$ and therefore the numerator of 4.3.i equals

$$\frac{\partial^2\Omega}{\partial p_i \partial p_j} = -\frac{\alpha}{2} < 0.$$

Case 2: If both firms follow the non-P strategy, then the negative slope of the reaction function is more accentuated:

$\Omega = [(1-p_j)\cdot\beta]\cdot p_i - r_i$ and therefore $\partial^2\Omega/\partial p_i\partial p_j = -\beta$.

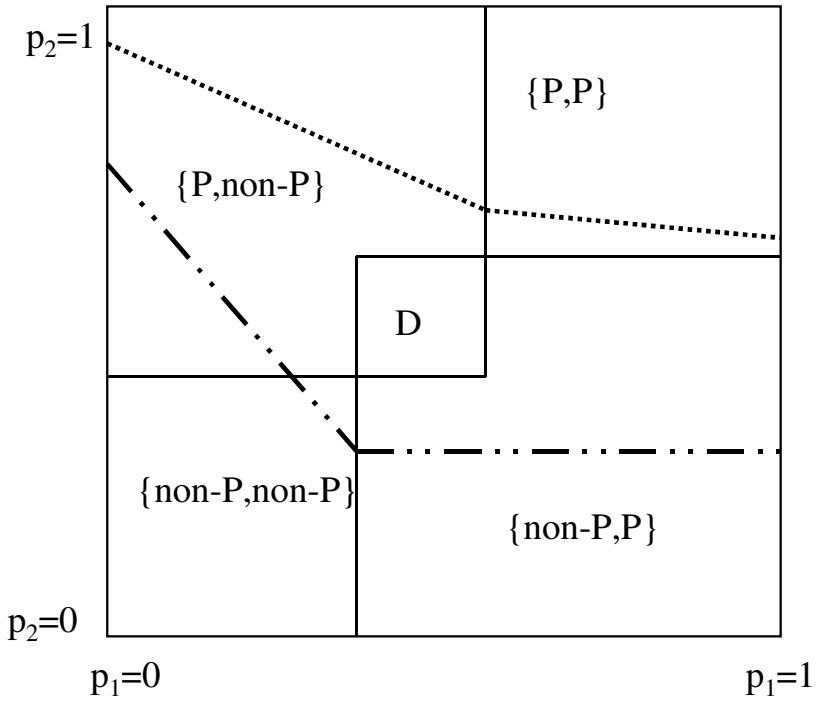
Case 3: If firm i follows the P strategy and firm j the non-P strategy, then in the case of risk neutrality $\Omega = p_i\alpha - r_i$, and therefore $\partial^2 U/\partial p_i \partial p_j = 0$.

Case 4: If firm i follows the non-P strategy and firm j the P strategy, then $\Omega = [(1-p_j)\cdot\beta]\cdot p_i - r_i$, and therefore $\partial^2\Omega/\partial p_i \partial p_j = -\beta$.

For a risk-averse firm, the reaction functions would have somewhat different slopes, but at least if the degree of risk aversion is moderate, the results derived in this section would not be entirely reversed.

Figure 13 shows a few possible firm 2 reaction functions. As can be seen, a firm that applies the P-strategy does not react at all to changes in competitive pressure caused by the rival, if the rival does not patent (case 3). By contrast a firm that does not patent its invention reacts strongly to competitive pressure (cases 2 and 4). If both firms apply the P-strategy, they react moderately to changes in competitive pressure (case 1).

Figure 13. **Two possible firm 2 reaction functions at the R&D stage. The optimal value of p_2 as a reaction to different values of p_1.**



Nothing guarantees there will be only one subgame perfect equilibrium at the R&D stage.

Firms can also have strategic motivations in respect of R&D decisions, because they may have incentives to affect rival patenting decisions and to commit themselves to certain policies. For instance, if firm 1 cannot count on achieving the favourable outcome {P, non-P} in the double equilibria area (D), it might restrict its own R&D investment. By restricting its possibilities to invent (p_1), firm 1 could discourage its rival from patenting, so that banks' reaction curves would cross at the patenting stage in the {P, non-P} area and not in the D area. This could in many cases be profitable for firm 1 even if it cannot limit its R&D expenditures by this decision. These kinds of strategic commitments will not be analysed in detail.

4.4 Predictions of the model and empirical observations

4.4.1 Previous findings

The model offers some potentially interesting new insights into the stylized facts presented above.

According to the first stylized fact, there should be a rather immediate industry-level correlation between R&D effort and patenting. This observation is consistent with result 4.3.c: a firm would react to increased rival R&D by patenting its inventions. If both firms belong to the same industry, this reaction will cause some positive correlation between R&D and patenting at the industry level.

According to the second stylized fact, R&D effort and patent count covariate in the long term, irrespective of the level of aggregation. This observed regularity is consistent with the model if we interpret the non-P strategy as postponed patenting, not non-patenting. If every invention is finally be patented, there will be a deterministic long-term causality running from innovative output to number of patent applications.

According to the third stylized fact, there is almost no immediate connection between number of patent applications filed by a firm and its own R&D effort. This is consistent with result 4.3.d: depending on the case, an increase in the innovative output of firm i either increases or decreases patenting by firm i itself. Moreover, if rival R&D is low, there is no connection between own R&D and patenting, because no patent applications will be filed anyway.

In addition, it is also possible to find some weak evidence in the previous literature that supports the hypothesis of countercyclicity of

patenting. Saint-Paul (1993, p. 881) calculated statistical covariations between aggregate demand shocks and two measures of technological efforts, namely R&D expenditures and national patent applications, using data from the major OECD countries. Nowhere but in the UK was patenting statistically significantly countercyclical but, on the other hand, none of the countries showed positive covariation with statistical significance. Saint-Paul reported no covariations between R&D expenditures and patenting.

4.4.2 Estimations using Finnish data

4.4.2.1 Patenting in Finland at the industry level

The hypothesis to be tested

The aim of the following analysis is to test one of the main empirical predictions of the model presented above: if a firm increases its R&D effort, the number of patent applications filed by its rivals will increase (result 4.3.c).

Finnish industry-level data are likely to have some properties that are of interest from the point of view of this hypothesized interrelationship. Finland is a small economy. *In the technology race, a typical Finnish firm faces mainly foreign competition.*

According to the model presented above, a firm becomes more willing to patent its inventions if its rival carries out a lot of R&D. This hypothesis has the following implication. When a Finnish firm intensifies its R&D efforts, its foreign rivals react by patenting a larger proportion of their inventions. This change is likely to affect their patenting behaviour both in Finland and in the rest of the world.

The model generates much weaker empirical predictions concerning the correlation between R&D carried out by Finnish firms and patent applications filed by domestic applicants, at least in the short run (result 4.3.d).

One might well question whether Finnish firms can affect foreign firms' global patenting behaviour because of the small size of the country. However, one should keep in mind that these foreign firms that would consider filing patent applications in Finland (most firms in the world would not) comprise a very biased sample. For them, Finland is often an important market, and a major source of competitive pressure. In respect of certain technologies related eg to ice-breakers and to peat as a source of energy, Finland may be the main source of a foreign firm's competitive pressure. If a firm with a

secret new product has good reason to presume that one of its rivals is working on the same idea, the inventor has an incentive to apply for a patent in the rival's country rather than in some other country. In addition, at least in the case of process inventions, it is difficult to understand why one would patent minor improvements in a certain country if there were no local manufacturers. Therefore, probably most foreign firms that apply for patents in Finland have Finnish competitors.

The data and the analysis

Statistics Finland has published estimates of the number of patent applications filed by domestic and foreign firms in different industries in 1980–1988.[15] These are given separately for more than 20 product groups. Product-group-specific data on R&D efforts of Finnish firms in 1981, 1983, 1985 and 1987 are presented in another publication of Statistics Finland.[16]

In the combined panel data, there are 24 product groups[17] and four observation years, ie 96 observations in all. In order to correct the R&D expenditures for inflation, the expenditures were divided by the wholesale price index. All the data are logarithmic. In the aircraft industry, there were two years in which no domestic patent applications were filed. In these cases, the logarithmic variable was somewhat arbitrarily coded as –2. However, in most other industries the number of patent applications was much higher, and the assumption that the errors are normally distributed is not completely unrealistic.

It was tested whether it would be possible to explain the number of domestic patent applications by domestic R&D effort in a simple

---

[15] Statistics Finland: Koulutus ja Tutkimus (Education and Research), 1989:24, p. 55 and p. 58.

[16] Virtaharju and Åkerblom: Technology Intensity of Finnish Manufacturing Industries, Helsinki, November 1993, p. 93.

[17] 1) Alimentary industries, beverages and tobacco, 2) Textile products, clothes, leather products and footwear, 3) Timber and wood products, 4) Pulp and paper, 5) Publishing and printing, 6) Furniture, 7) Chemicals, 8) Pharmaceuticals, 9) Other chemical products, 10) Rubber and plastic products, 11) Clay and stone products, 12) Iron, steel and other basic metals, 13) Metal products, 14) Pulp and paper making machines, 15) Metallurgical & other machinery, 16) Electrical equipment, 17) Radio, TV and telecommunication equipment, 18) Computers and office machines, 19) Instruments and optical equipment, 20) Ships, 21) Aircraft, 22) Automobiles and other transport equipment, 23) Energy and water supply, 24) Construction.

dynamic structure. The number of patent applications was regressed on its own value lagged by two years, current domestic R&D expenditure, and domestic R&D expenditure lagged by two years. As in the model, it was assumed that R&D effort affects patent application, but not *vice versa*.

Both the fixed-effects model and the random-effects model were tested. In terms of the Hausman test, the random-effects model did not fit the data at all, and it was clearly rejected. In the two-way fixed-effects model, year-specific as well as industry-specific effects were controlled for with dummy variables. The explanatory variables include even a general constant.[18]

Table 4.1 **Patenting by Finnish firms in Finland as a function of Finnish R&D efforts, fixed-effects panel data estimation results for years 1983, 1985 and 1987**

Least Squares with Group Dummy Variables and Period Effects
Ordinary least squares regression.

Dep. Variable	=	LDOPA
Observations	=	72
R-squared	=	0.926
Adjusted R-squared	=	0.875
F[29, 42]	=	0.181 E+02
Estd. Autocorrelation of e(i,t)		–0.199
Mean of LDOPA		2.843
Std dev of LDOPA		1.405
Mean of LREEX		8.464
Std dev of LREEX		1.251

Variable	Coefficient	Std Error	t-ratio	Prob |t| x
LDOPA[–2]	–0.749	0.163	–4.59	0.000***
LREEX	0.633E–01	0.330	0.19	0.849
LREEX[–2]	0.322	0.295	1.09	0.278
Constant	1.794	3.831	0.47	0.641

LDOPA = Ln (patent applications filed by domestic applicants)
LREEX = Ln (Real R&D expenditure by Finnish firms)

As one can see in the table 4.1, this analysis provides very little if any evidence that domestic R&D and patenting by Finnish firms are

[18] This constant is a linear combination of year-specific and industry-specific dummy variables, but the problem of multicollinearity is avoided by imposing on the dummy variables the restriction that the sum of their coefficients must equal zero.

somehow associated in the short run. If such covariation exists, it must be weak. This finding is consistent with result 4.3.d.

These results may seem to be inconsistent with the findings of Pakes and Griliches (1984) using US data. Hence there may be a fundamental difference between the two countries in patenting. In the light of the model presented above, such a difference would be understandable. An obvious difference between Finland and the US is the size of the country. In the US, many firms have domestic competitors rather than foreign ones. If a firm increases its R&D effort, its rival files more patent applications. Both phenomena are reflected in domestic statistics, because competition is mainly domestic. If a Finnish firm increases its R&D effort, there may be a comparable increase in rival patenting activity, but the increase takes place abroad and is not reflected in Finnish patent counts.

In Finland the number of patent applications seems to have a very pronounced tendency to fluctuate over time: a large number of patent applications in one year seems to be a good indicator that there will be relatively few applications in the next year. This effect is quite strong and highly significant. It could mean that firms sometimes prefer to postpone applications, which reduces the number of applications now but increases them in the future.

It was then tested with a comparable specification whether there is any connection between the number of foreigners' patent applications and R&D efforts of Finnish firms in the respective product groups.

Table 4.2 **Patenting by foreign firms in Finland by product group as a function of Finnish R&D efforts, fixed-effects panel data estimation results for years 1983, 1985 and 1987**

Least Squares with Group Dummy Variables and Period Effects
Ordinary least squares regression.

Dep Variable	=	LFOPA
Observations	=	72
R-squared	=	0.993 + 00
Adjusted R-squared	=	0.988 + 00
F[29, 42]	=	0.210 + E03
Estd Autocorrelation of e(i,t)		–0.234
Mean of LFOPA		4.103
Std dev of LFOPA		1.366
Mean of LREEX		8.464
Std dev of LREEX		1.251

Variable	Coefficient	Std Error	t-ratio	Prob |t| ≥ x
LFOPA[–2]	–0.133	0.872E–01	–1.53	0.13148
LREEX	0.226	0.941E–01	2.40	0.01901*
LREEX[–2]	0.927E–01	0.849E–01	1.09	0.27863
Constant	2.074	1.114	1.86	0.06696

LFOPA = Ln (Patent applications filed in Finland by foreign applicants)
LREEX = Ln (Real R&D expenditure by Finnish firms)

As we see from table 4.2, there is a statistically significant immediate covariation between Finnish R&D and patenting by foreign firms in Finland. This is consistent with result 4.3.c of the model: patenting by a firm is correlated with the rival's R&D effort.

It was found that there was a problem concerning autocorrelation in the residuals of both regressions presented above. In both cases, the autocorrelations are negative and close to the borderline of statistical significance at the two-tailed 5 % level with 72 observations. In the case of foreign applications, the autocorrelation exceeds the borderline, and in the case of domestic applications, it is just below it. However, correcting for autocorrelation with the AR1 would consume one degree of freedom per industry. Because the data are of limited temporal dimension, this is extremely costly. In testing, the AR1 correction, it was found that the estimated regression coefficients did not undergo major changes.

It is possible that the number of foreign patent applications is related to Finnish R&D expenditure only because R&D efforts in Finland are strongly correlated with R&D efforts abroad and because

foreign R&D has a positive impact on the number of foreign patent applications filed in Finland. There apparently are no data available on the amount of R&D expenditure that have been systematically collected from firms operating in different countries, with different firms being given weights according to the intensity of technological competition between them and their Finnish competitors. Thus this potential explanation cannot be adequately tested. However, this hypothesis would be inconsistent with the observation that there is virtually zero covariance between domestic patenting and domestic R&D.

To sum up, these results are inconsistent with the conventional view that firms file more patent applications if they invest heavily in R&D. Instead, they are consistent with the model presented above: more R&D by a firm induces its rival to file more patent applications.

4.4.2.2 Is aggregate patenting counter cyclical?

In this section, we look at whether the trend-deviations of patent counts and real GDP are negatively correlated, as one might expect in the light of the model presented above.

In macroeconomics, it has become fairly common to study the relationship between inflation and real GDP via the correlation between the trend-deviations of these two variables. Kydland and Prescott (1990) is a seminal article in this regard. In the following, a similar method will be used to analyse the cyclicity of patent counts.

The yearly trend values for logarithmic real GDP and patent counts over the period 1954–1995 were estimated using the Hodrick–Prescott filter.[19] The difference between actual value and trend value was calculated for every year for both variables.

---

[19] Trend values for the variable x for the periods 1, 2, ..., T are defined as the time series that minimizes the value of the following function.

$$\sum_{t=0}^{T} \left(x_t - y_t\right)^2 + \lambda \sum_{t=1}^{T-1} \left[\left(y_{t+1} - y_t\right) - \left(y_t - y_{t-1}\right)\right]^2$$

where x_t = original non-filtered observation for period t, y_t = calculated trend value for the same variable and observation period, and $\lambda>0$ is an arbitrarily chosen exogenous constant.

Strictly speaking this method as a statistical tool was not developed by Hodrick and Prescott, but according to Kydland and Prescott (1990), they were the first researchers known to have applied it in economics.

In the following table, we can see how the trend deviation of patent counts is correlated with various leads and lags of the trend deviation of real GDP.

Table 4.3 **Correlations between trend deviations for real GDP (incl. leads and lags) and number of domestic patent applications in Finland.**

	$\lambda = 10$	$\lambda = 100$	$\lambda = 1\,000$	$\lambda = 10\,000$
GDP+4	–0.095	0.041	0.074	–0.011
GDP+3	0.032	–0.023	0.007	–0.171
GDP+2	0.193	–0.063	–0.049	–0.140
GDP+1 (Lead)	–0.234	–0.370	–0.351	–0.235
GDP	–0.537	–0.556	–0.522	–0.320
GDP–1	–0.418	–0.419	–0.405	–0.237
GDP–2	0.211	0.142	0.125	0.215
GDP–3	0.429	0.460	0.448	0.504
GDP–4	0.273	0.484	0.491	0.538

Data covers the years 1954–1995. If GDP is lagged for n years, the patent counts are for the period (1954+n)–(1995). For correlation between trend deviation of patent count and the n:th lead of GDP, the patent counts are for years (1954)–(1995-n).

As we can see, there is a relatively strong immediate negative correlation between trend deviations of GDP and patent counts. This is consistent with the model, provided we assume firms to be decreasingly risk averse. Because the trend-deviations are strongly autocorrelated, one cannot rely on any simple measure of the degree of statistical significance.

Interestingly, the third and fourth lags of GDP trend-deviation are clearly positively correlated with the trend-deviation of patent counts. If we assume that instead of not filing patent applications, firms simply postpone their applications during booms, this finding accords with expectations. The immediate impact of a boom on patenting is negative, but because the applications are filed at a later date, the lagged impact is positive.

One interesting question has to do with the composition of the flow of patent applications. Applications are filed by both firms and private individuals. Is the observed countercyclicity of patent applications due to the firms' behaviour or to the behaviour of private inventors? In 1990 at the start of the deep recession of the early 1990s, patent applications filed by domestic business firms increased by

about 12 % from the previous year, whereas the number of applications filed by private individuals remained almost constant.[20] At least in this case the increase in patent applications that took place simultaneously with an economic slowdown was caused by the business sector.

The model presented above provides a possible explanation for the relative strength of the counter-cyclicity of patenting that seems to prevail in Finland as compared to many other countries. Business cycles in different countries are imperfectly correlated. Because patenting decisions are strategic substitutes in the model, the positive impact of a recession on patenting is stronger if the rival does not react to the same macroeconomic shocks as the firm in question. If anything, a recession that affects the rival makes a firm less willing to patent (result 4.3.f). Hence, the smaller the country, the stronger the counter-cyclicity of patenting is likely to be, because in small countries, the competitive pressure is predominantly of foreign origin.

4.4.2.3 A few comments on the time series properties of the variables

This section discusses the time series properties GDP volume and the number of patent applications filed by domestic applicants. Both variables are measured as logarithmic annual observations for the period 1954–1996.

The concept of unit root processes has become central in time series econometrics. The Dickey-Fuller unit root test results were –1.142 for the logarithmic GDP volume and 0.3972 for logarithmic patent applications, when no lags or trend were used. In terms of the statistical significance tests presented by MacKinnon (1991), these values are not statistically significant at the 5 % level, and the null hypothesis of a unit root cannot be rejected.

It might also be of some interest to know whether there are any cointegrating vectors between the two variables. By applying the Johansen multivector model with two lags, it is possible to find two eventual cointegrating vectors. The trace values are 1.895 for the null hypothesis of less than two vectors, and 17.83 for the null hypothesis of no vectors. Johansen and Juselius (1990) presented statistical significance level tables for trace tests. In the light of these tables, it is

[20] Statistics Finland, koulutus ja tutkimus (Education and Research), 1992:2 ’Teknologian soveltaminen ja siirto’.

not possible to reject the null hypothesis of no cointegrating vectors at the 5 % significance level.

At the 10 % significance level, the "stronger" candidate could be accepted. By normalizing the patent counts, the beta coefficient of Ln(PAT) in the cointegrating vector is set to equal +1, and the coefficient of Ln(GDP) is –0.998. This implies that a 1 % increase in GDP would in the long run be associated with a 0.998 % ≈ 1 % increase in the number of patent applications.

The adjustment parameters (alphas) reflect how strongly the two variables would react to deviations from the cointegrating relationship. The results imply that it is patenting rather than GDP that reacts. The alpha for Ln(PAT) is –0.234, whereas the alpha for Ln(GDP) is +0.035. Thus, if the number of patent applications is disproportionately low in comparison to GDP, the number of patent applications will have a strong tendency to increase, but it is unlikely that GDP would decline as a result.

4.5 Discussion of the model

The traditional view concerning patents and R&D has been fairly simplistic. Firms have been assumed to file patent applications whenever they have innovative output to be patented. This view is not entirely consistent with empirical observations. Nevertheless, surprisingly few theoretical contributions have analysed the "to patent" dilemma.

The model presented above is based on the idea that patenting is a kind of insurance. An invention often enables the innovator to earn monopoly profits. Because patent applications are public documents, patenting helps rivals and imitators, which erodes profits. On the other hand, by not patenting the innovative firm takes the risk that it will earn no monopoly profits at all with its invention, because a rival might invent the same product or process. Thus, a firm is likely to patent its inventions if it either faces strong competitive pressures or is strongly risk averse.

The model finds at least preliminary support in Finnish data. Foreign firms have had a strong propensity to patent in Finland if their Finnish competitors have invested heavily in R&D. In contrast, there seems to be no immediate correlation between Finnish R&D and patenting by Finnish firms. The number of domestic patent applications has been systematically higher during recessions than during booms, which suggests that the propensity to patent may have

been higher during recessions. A possible explanation for the countercyclicity of patenting is that innovators may have been decreasingly risk averse. Because there seems to be no covariation between domestic patent counts and domestic R&D effort in the short run, this observation should not be regarded as evidence against real business cycle theories. The number of patent applications and GDP seem to be positively related in the long run.

Appendix 1

Result 4.3.a
It is not possible to construct examples where there is no subgame perfect equilibrium at all at the patenting stage.

Proof:

There is no subgame perfect equilibrium if

either (possibility 1)
a) {P, Non-P} is not an equilibrium because firm 2 would react to P by firm 1 with P,
b) {P, P} is not an equilibrium because firm 1 would react to P by firm 2 with Non-P,
c) {Non-P, P} is not an equilibrium because firm 2 would react to non-P by firm 1 with non-P,
d) {Non-P, Non-P} is not an equilibrium because firm 1 would react to non-P by firm 2 with P.

or (possibility 2)
– {P, P} is not an equilibrium because firm 2 would react to P by firm 1 with non-P,
– {P, Non-P} is not an equilibrium because etc., as above.

The first possibility can be ruled out.

If the conditions for the first possibility are satisfied, the payoff functions for firm 2 satisfy the following conditions: (4.3.a.*) and (4.3.a.**).

Stage a) implies that if firm 1 applies P, firm 2 prefers P

$$\Rightarrow \tfrac{1}{2}p_i\alpha + (1-p_1)\alpha > (1-p_1)\beta \qquad (4.3.a.*)$$

and the stage c) implies that if firm 1 applies non-P, then firm 2 prefers non-P

$$\Rightarrow \alpha < (1-p_1)\beta. \qquad (4.3.a.**)$$

Conditions (4.3.a.*) and (4.3.a.**) imply

$$\alpha < (1 - p_1) \quad \beta < ½p_1\alpha + (1 - p_1)\alpha$$
$$\Rightarrow \alpha < ½p_1\alpha + (1 - p_1)\alpha$$
$$\Leftrightarrow 0 < ½p_1 - p_1,$$

which is impossible.

The second possibility can be ruled out in an analogous way.

QED

Appendix 2

Result 4.3.b
If either {P by both firms} or {non-P by both firms} is a subgame perfect equilibrium, there cannot be any other subgame perfect equilibria.

Proof:

If {P, P} is an equilibrium, P is the dominant strategy for both firms (see footnote 13) ⇒ No other outcome could be a subgame perfect equilibrium.

{non-P, non-P} cannot be an equilibrium unless

$$p_2 < \frac{(\beta - \alpha)}{\beta};$$

$$p_1 < \frac{(\beta - \alpha)}{\beta}.$$

An equilibrium where either of the firms plays P while its rival plays non-P cannot prevail unless either of the two firms has a probability to innovate higher than $(\beta-\alpha)/\beta$ which is inconsistent with $p_2 < (\beta-\alpha)/\beta$; $p_1 < (\beta-\alpha)/\beta$.

It is not possible that both {non-P, non-P} and {P, P} would be subgame perfect equilibria simultaneously, because if {P, P} is an equilibrium, it is the dominant one.

⇒ If {non-P, non-P} is a subgame perfect equilibrium, it is unique.

QED

5 Conclusions

5.1 Main findings and policy implications

Firms' incentives to develop new products and processes have been analysed in various branches of economics at least since the early 19th century. The analysis presented in the present study belongs to a rapidly expanding research tradition that emerged about two decades ago, namely game theoretic oligopoly models. Hopefully, it provides us with new insights on certain specific topics related to technology competition between companies in duopolistic industries.

In the introduction, the history of endogenous innovations in economics was briefly reviewed. Moreover, modern economic literature concerning innovation policies and banking technology was discussed in closer detail, as well as some of the literature on patenting, because these topics were of particular relevance to the models presented here.

The second chapter of this study presented a simple model of duopolistic bank competition. Unlike in many previous microeconomic contributions concerning oligopolistic competition in banking, the emphasis here was on payment services rather than borrowing and lending in the presence of asymmetric information. Payment services as a competitive tool in the struggle for market share has not been a central topic of previous theoretical research. In this sense, chapter 2 presents a pioneering model. Banks' incentives to invest in developing the payment system were the main topic of analysis. The quality of interbank payment services offered to customers was determined endogenously by the investment decisions of the two banks.

One of the main finding of this model is that banks often have distorted incentives to develop the interbank payment system if they either cannot price the payment service or if they voluntarily prefer not to charge fees for payment services. If financial market regulation or insufficient interest rate competition maintains abnormally wide interest rate margins, thus making lending and deposit taking unnaturally profitable for banks, attracting more customers becomes the banks' main objective. Banks can offer free payment services to attract more customers. If payment services are not a source of income but rather a marketing tool, a large bank will have an incentive to deter the smooth functioning of the interbank payment system. If bank A has a 90 % market share, it is unattractive to be a bank B customer

unless interbank payment transactions are at least satisfactory. If payment transactions with the customers of the large bank (A) take weeks and are sometimes not possible at all, the small bank (B) might be unable to get any customers. Thus the dominant bank (A) may try to increase its market share further at the cost of its small rival by not investing in the interbank system. The small bank (B), would instead prefer to invest excessively in order to alleviate the problems caused by its small market share.

If banks do charge a fee for using the payment system, their incentives are less distorted. Investing in the system helps even a large bank to earn more fee revenue.

Because insufficient interest rate competition may indirectly distort the allocation of resources in a payment system, antitrust policies against collusion in respect of loan and deposit services can have beneficial indirect effects on the allocation of resources in the payment system.

However, if banks charge fees for making payments, they often overinvest in the system. The over-investment is partly due to the fact that a well functioning system would paradoxically relax price competition, which would obviously be in the banks' interest. The intuition behind this surprising result is quite simple. If interbank payments are slow and unreliable, it is useful for customers to use the same bank as the majority of other customers. Thus, if a bank increases its market share by cutting its prices, the additional customers attracted by low fees will in turn attract even more customers who might have to exchange payments with them. If interbank payment services are highly developed, the cumulative effects of gaining an additional customer are weak, because the possibilities to exchange payments would not depend as strongly on banks' market shares. Therefore, the better interbank payments function, the less the indirect cumulative effects. Thus, with a good interbank payment system, market shares react much less strongly to banks' pricing decisions. This obviously lessens price competition.

The model was also used to analyse the incentives of a welfare maximizing central bank to investment in the interbank system. The central bank can affect the quality of the payment system in two different ways. First, such investments affect the quality of the system directly. Secondly, these investments affect the behaviour of private banks. The findings imply that central bank involvement is particularly essential if two banks of roughly equal size offer payment services free of charge, because private investment in the system is sub-optimal. If, instead, two banks of equal size charge fees for using the payment system, central bank involvement is less essential, and

there is little need to try to affect the behaviour of private banks. Hence the optimal role of the central bank depends on banks' pricing policies.

Chapter 3 presents a model on discriminatory patent protection. It seems that in real life national governments and patent officials have sometimes favoured domestic firms either in patent legislation or in administrative practice. Aoki and Prusa presented a pioneering theoretical analysis of the impact of such policies on firms' R&D efforts. Because discriminatory policies have been a real-life phenomenon, their contribution deserves to be extended.

Chapter 3 extends their basic model in two different ways. In the first version, it is assumed that there are two companies based in different countries. If the two domestic governments discriminate against each other's firms in order to favour domestic companies, these protectionist policies may offset each other's effects. At the firm level, being favoured in the home country and discriminated against in the foreign country can have equally strong but opposing effects, and therefore discrimination may have no impact on firms' R&D efforts. Secondly, it was demonstrated that unilateral discriminatory protection offered by the government to a domestic company competing against a foreign rival may either encourage or discourage domestic R&D, depending on what kinds of incentives the domestic company has to intensify or maintain its R&D investment. If intensifying the R&D effort is useful mainly because additional expenditure increases the likely value of the invention, discrimination by the home country government would mainly decrease the main risk related to the investment, namely the possibility that the rival would win the patent race. Obviously, reducing the risk of the investment would make it more attractive to spend more money on the project. If, instead, additional R&D mainly increased the probability of getting the patent but had a minor impact at most on the value of the patented invention, protectionist policies by the domestic government would have an entirely different effect. Expensive R&D efforts by the firm could be replaced by such protectionist policies. The firm could simply save money by reducing its R&D efforts without jeopardizing its possibilities to get the patent.

Though chapter 3 probably has few robust policy implications, it may be of interest because it questions conventional wisdom. The standard analysis concerning the benefits of free trade is well known and it seems to have affected government policies in different parts of the world. The standard analysis is particularly useful in describing and predicting the likely consequences of tariff protection imposed on foreign imports in competitive industries. Discrimination against

foreigners in the case of intellectual property rights in oligopolistic industries has little to do with such tariff protection, and it may have entirely different effects. It is far from obvious that international agreements aimed at dismantling discriminatory patent legislation contribute to global welfare. At least it was demonstrated that the welfare effects of discriminatory patent protection are far less straightforward than the effects of tariff protection under perfect competition.

Chapter 4 focuses on firms' incentives to patent their innovations in order to analyse whether patent statistics are a satisfactory indicator of inventive output. The innovative output itself is, in practice, exogenously given in the model, and the main emphasis is on the "to patent" dilemma.

The model has been strongly inspired by certain empirical observations presented in previous literature. It has been observed that the amount of resources spent on research and development and the number of patented inventions are fairly strongly correlated at the industry level, whereas the firm-level correlation is substantially weaker, if it is observed at all. The short-term relationship between patents and innovative effort seems to have been particularly weak at the company level.

In the model, patenting is a kind of insurance. By patenting its invention a company does not achieve the highest possible level of profits. Instead, patenting reduces risks. An invention enables the innovator to earn high profits if and only if the invention is monopolized. The innovator can maintain its monopoly position if it either patents the innovation or manages to keep essential details secret, thus preventing competitors from imitating. Because patent applications are public documents, patenting helps rivals and imitators, which erodes profits. In real life, patents do not guarantee perfect monopoly rights. Thus, secret innovations are more valuable than patented ones. On the negative side, by not patenting the innovative firm runs the risk that it will earn no monopoly profits at all with its invention because a rival might invent the same product or process. Thus a firm is likely to patent its inventions if it either faces strong competitive pressures in the technology race or is highly risk averse. These predictions are consistent with previous empirical findings.

The model receives further support in panel data on Finnish manufacturing industries in the 1980s. Foreign firms have had a higher than average propensity to patent in Finland when their Finnish competitors have invested heavily in R&D. On the other hand, there seems to be no immediate industry-level correlation between Finnish

R&D and patenting by Finnish firms. In the light of macro-level data, during recessions the number of domestic patent applications has been systematically higher than during booms; this implies that the propensity to patent may have been higher during recessions. Assuming innovators have been decreasingly risk averse, the countercyclicity of patenting can be explained by the hypothesis that patenting is a kind of insurance.

Chapter 4 has few policy implications, except at the meta level. In any policy-making, it is essential to base decisions on valid and reliable statistics and indicators concerning important economic and societal phenomena. In the light of the results, the relevance of patent counts as a technology indicator is highly questionable. This indicator may be particularly biased in a small country where companies typically face foreign rather than domestic technology competition. The patenting propensity of domestic companies might reflect the R&D activities of their foreign rivals rather than domestic investment in the development of new technologies.

Chapters 2 and 3 implicitly emphasize the fact that the government can try to encourage or deter firms' investments in new technologies without taxing or subsidizing R&D. In real life, direct subsidies and lending on favourable terms are often used as government innovation policy tools and as industrial policy tools in general. This might reflect certain historical factors. In the past, when financial markets were still regulated in many countries, firms were often financially constrained. Thus, the possibility to get loans as a reward to a certain decision probably affected firms' behaviour quite strongly. Financial markets are no longer regulated and financial incentives have probably lost part of their traditional effectiveness. Public funding of private research is by no means the only feasible policy tool.

5.2 Suggestions for further research

Chapters 2–4 have hopefully provided some new insights into certain topics related to investment in technological progress. These approaches could be elaborated by taking into account, for instance, the following factors.

The models are based on the basic assumption that the industry is a duopoly. It is not clear whether the main policy implications would be robust if the number of companies were larger, or if the number of companies were endogenized. For instance, in the case of payment systems, in most countries there is no dominant bank with a higher-

than-50 % market share. Moreover, it might be of some interest to analyse structures with a large dominant company competing against a competitive fringe consisting of a large number of small firms.

To take a concrete example of effects that are neglected by assuming the duopoly structure, one could mention alliances between companies within an industry. R&D joint ventures between innovative firms have become a topical issue in the economics literature, probably because these kinds of alliances have been observed in the real life. It is not possible to analyse with duopoly models optimal government policy responses if two firms in an industry form such an alliance in order to compete against other companies. It might be of interest to find new arguments either in favour of government policies aimed at preventing such collusive behaviour or *vice versa*.

This point might be of particular relevance in the case of payment systems. Certain bank groups cooperate, and they have created mutual networks to facilitate payment traffic among themselves. It has been observed that small banks cooperate particularly often in various activities, including payment systems, and it might be fruitful to analyse whether an extended version of the payment system model would predict the emergence of such alliances. Moreover, the model might have policy implications concerning optimal central bank policies toward such alliances.

As a second point one could consider the dynamic nature of technology. The progress of technology is a process; by excluding the dynamic nature of the phenomenon one runs the risk that important issues are neglected. Innovations are often based on previous innovations, and because of learning-by-doing effects, accumulation of knowledge and other such reasons, companies that have made successful inventions in the past may have a competitive advantage when the next generation of technologies is to be developed. Because the models are rather static, these aspects have been largely neglected in this study. Perhaps some of the models could be developed into dynamic versions, or possibly some of the approaches could be presented as supergames, which might eventually reveal different effects than those presented in this work.

As a third suggestion one might propose that the possible heterogeneity of firms be taken into account. Developed and developing countries can have conflicts of interest in the case of intellectual property rights. The models concerning patenting could be extended by assuming that the competition takes place between a company based in a developed country and its rival based in an LDC. The LDC-based company would have some exogenously imposed limitations in its capability to create innovations. For instance, one

could assume that the LDC-based company has no possibility of getting patents unless its domestic government favours it.

In addition, the two firms might differ in the following ways. Firms' degree of risk aversion and other factors related to their objective functions may differ. The models could be combined with theories concerning, say, mixed oligopolies. Because public ownership has been relatively common in banking, it might be particularly meaningful to assume in the context of chapter 2 that the banking industry is a mixed duopoly. Moreover, in any of the models either of the two firms could be a managerialistic firm without effective ownership control, aiming at maximal prestige for its managers by maximizing turnover. Or, to take still another example, one might assume that either of the two firms would be a Stackelberg leader.

As concluded in the introduction, when compared to the total number of different papers published concerning these issues, there have been relatively few contributions where game theoretic analyses is combined with empirical work. This study, notably chapter 4, has hopefully contributed to narrowing the gap. Even though the predictions of the payment system model presented in chapter 2 are not strictly speaking tested, certain analogies between the predictions of the model and the history of Finnish payment systems are found. In principle, it might be possible to collect at least some data concerning, say, banks' investments in payment systems in different countries at a given moment of time. Correlations between the propensity of banks to make such investments, banks' pricing policies and their market shares might provide us with interesting further evidence concerning the hypotheses presented in the second part of this study.

On the other hand, the main conclusions of chapter 3 may be more difficult to test empirically; the impact of patent legislation on firms' R&D efforts is difficult or even impossible to measure. In this sense, the second model has the same limitations as many previous game theoretic oligopoly models; the main conclusions are not directly testable.

Despite all the limitations of this study, it is hoped that it has both contributed a few innovative theoretical insights and accomplished at least a minor narrowing of the gap between theory and empirics.

References

Aaku, E. (1957) **Suomen liikepankit 1862–1955.** Unpublished manuscript.

Acs, Z.J. – Audretsch, D.B. (1989) **Patents as a Measure of Innovative Activity.** Kyklos 42, 171–180.

Adams, L.A. (1998) **International Standards for IP Protection and R&D Incentives Revisited.** Open Economies Review 9:4; 343–348.

Aghion, P. – Howitt, P. (1992) **A Model of Growth Through Creative Destruction.** Econometrica 60, 323–351.

Alhonsuo, S. – Tarkka, J. (1989) **Rahoitustoiminnan tuottavuus ja tehokkuus Suomessa.** (The productivity and efficiency of financing activities in Finland) In Pankkitoiminnan kannattavuus, tehokkuus ja riskit Suomessa; Suomen Pankki A:72.

Altunbas, Y. – Molyneux, P. (1996) **Economies of Scale and Scope in European Banking.** Applied Economics 6, 367–375.

Angelini, P. – Giannini, C. (1994) **On the Economics of Interbank Payment Systems.** Economic Notes by Monte dei Paschi di Siena, vol 23, no 2, 194–215.

Aoki, R. – Prusa, T.J. (1993) **International Standards for Intellectual Property Protection and R&D Incentives.** Journal of International Economics 35, 251–273.

Arrow, K. (1963) **Economic Welfare and the Allocation of Resources for Invention.** The Rate and Direction of Inventive Activity, Princeton University Press.

Artus, P. – Kaabi, M. (1993) **Dépenses Publiques, Progrès Technique et Croissance.** Revue économique, 44, 287–317.

Auer, J. (1964) **Hyvinvoinnin rakennuspuita – Postisäästöpankki vuosina 1886–1961.** (Elements of welfare – the Postal Savings Bank in 1886–1961), Valtioneuvoston kirjapaino ja sitomo.

Badulescu, P. (1988) **R&D and Productivity Growth in Swedish Manufacturing industry 1963–1981. Models, Problems, Results.** Uppsala University, Department of Economics, Working paper 7/88.

Beard, T.R. – Caudill, S.B. – Gropper, D. (1997) **The Diffusion of Production Processes in the U.S. Banking Industry.** A Finite Mixture Approach, Journal of Banking and Finance 21, 721–740.

Beath, J. – Katsoulakos, Y. – Ulph, D. (1989a) **The Game-Theoretic Analysis of Innovation: a Survey.** Bulletin of Economic Research 41, 163–184.

Beath, J. – Katsoulakos, Y. – Ulph, D. (1989b) **Strategic R&D Policy.** The Economic Journal, Conference 99, 74–83.

Berg, B. – Kilvits, K. – Tombak, M. (1996) **Technology Policy for Improving Competitiveness of Estonian Industries.** ETLA C 73, Helsinki.

Berger, A. (1985) **The Economics of Electronic Funds Transfer.** Working Paper, Board of Governors of the Federal Reserve System.

Blackburn, K. – Hung, V.T.Y. (1993) **Endogenous Growth and Trade Liberalization.** Centre for International Economics Memo 5/93, University of Aarhus.

Bondt, R. de – Slaets, P. – Cassiman, B. (1992) **The Degree of Spillovers and the Number of Rivals for Maximum Effective R&D.** International Journal of Industrial Organization 10, 35–54.

Bouckaert, J. – Degryse, H. (1995) **Phonebanking.** European Economic Review 39, 229–244.

Bound, J. – Cummins, C. – Griliches, Z. – Hall, B.H. – Jaffe, A. (1984) **Who does R&D and who Patents?** In Griliches, Z.: R&D, Patents and Productivity, The University of Chicago Press.

Brander, J.A. – Spencer, B.J. (1983) **International R&D Rivalry and Industrial Strategy.** Review of Economic Studies 50, 707–722.

Bregman, A. – Fuss, M. – Regev, H. (1991) **High Tech and Productivity.** Evidence from Israeli Industrial Firms, European Economic Review 35, 1199–1221.

Choi, J.P. (1990) **Market Structure, Incentive to Patent and the pace of Innovation.** Economics Letters 34, 277–283.

Cincera, M. (1997) **Patents, R&D and Technological Spillovers at the Firm Level: Some Evidence from Econometric Count Models for Panel Data.** Journal of Applied Econometrics, 3/12, 265–280.

Cohen, W.M. – Klepper, S. (1992) **The Anatomy of Industry R&D Intensity Distributions.** American Economic Review 82, 773–799.

Cohen, W. – Levinthal, D. (1989) **Innovation and Learning, the Two Faces of R&D.** Economic Journal 99, 569–596.

Comanor, W.S. – Scherer, F.M. (1969) **Patent Statistics as a Measure of Technical Change.** Journal of Political Economy 77, 392–398.

Connolly, R.A. – Hirschey, M. (1988) **Market Value and Patents.** Economics Letters 27, 83–87.

Crépon, B. – Duguet, E. (1997) **Estimating the Innovation Function from Patent Numbers: GMM on Count Panel Data.** Journal of Applied Econometrics 3/12, 243–263.

Culbertson, J.D. (1985) **Econometric Tests of the Market Structural Determinants of R&D Investments: Consistency of Absolute and Relative Firm Size Models.** The Journal of Industrial Economics, 34, 101–107.

Dasgupta, P. – Stiglitz, J. (1980) **Uncertainty, Industrial Structure and the Speed of R&D.** Bell Journal of Economics 11, 1–28.

David, P.A. – Olsen, T.E. (1992) **Technology Adoption, Learning Spillovers and the Optimal Duration of Patent-Based Monopolies.** International Journal of Industrial Organization 10, 517–543.

De Fraja, G. (1993) **Strategic Spillovers in Patent Races.** International Journal of Industrial Organization 11, 139–146.

Degryse, H. (1996) **On the Interaction Between Vertical and Horizontal Product Differentiation: An Application to Banking.** The Journal of Industrial Economics XLIV, 169–185.

Delbono, F. – Denicolò, V. (1993) **Regulating Innovative Activity: The Role of the Public Firm.** International Journal of Industrial Organization 11, 35–48.

Denison, E.F. (1985) **Trends in American Growth.** Washington, DC.

Deolalilkar, A.B. – Röller, L.H. (1989) **Patenting by Manufacturing Firms in India: Its Production and Impact.** The Journal of Industrial Economics, March 37, 303–314.

Devinney, T.M. (1993) **How Well do Patents Measure New Product Activity?** Economics Letters 41, 447–450.

Ekelund, R.B. – Hébert, R.F. (1983) **A History of Economic Theory and Method.** Second edition.

EMI (1996) **Payment Systems in the European Union.** European Monetary Institute.

Englander, A.S. – Evenson, R.E. – Hanazaki, M. (1988) **R&D, Innovation and the Total Productivity Slowdown.** OECD Economic Studies 11, 8–42.

Entorf, H. (1988) **Die Endogene Innovation, Eine Mikro-Empirische Analyse von Produktphasen als Innovationsindikatoren.** Jahrbücher für National-ökonomie und Statistik 204, 175–189.

Ergas, H. (1987) **The Importance of Technology Policy.** In Dasgupta – Stoneman (eds): Economic Policy and Technological Performance, Cambridge University Press.

Esho, N. – Sharpe, I.G. (1995) **Long-Run Estimates of Technological Change and Scale Economies in a Dynamic Framework. Australian Permanent Building Societies 1974–1990.** Journal of Banking and Finance 19, 1137–1157.

Evenson, R.E. (1993) **Patents, R&D and Invention Potential – International Evidence.** American Economic Review 83, 453–468.

Evenson, A.E. – Steven, R. – Hanazaki, M. (1988) **R&D, Innovation and the Total Factor Productivity Slowdown.** OECD Economic Studies, Autumn, 8–42.

Ferrantino, M.J. (1992) **Technology Expenditures, Factor Intensity and Efficiency in Indian Manufacturing.** The Review of Economics and Statistics 74.

Fisher, F. (1989) **Games Economists Play: A Noncooperative View.** Rand Journal of Economics 20, 113–124.

von Franke, J.F. (1993) **Die Bedeutung des Patentwesens im Innovationsprozeβ.** Ifo Studien 3–4:39, 207–326.

Fransman, M. (1995) **Is National Technology Policy Obsolete in a Globalised World? The Japanese Response.** The Cambridge Journal of Economics 19, 95–119.

Freeman, C. (1987) **Technology Policy and Economic Performance.** Lessons from Japan, Pinter Publishers.

Frei, F.X. – Harker, P.T. – Hunter, L.W. (1998) **Innovation in Retail Banking.** The Wharton School, University of Pennsylvania; 97-48-B.

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

Fölster, S. (1991) **The Art of Encouraging Innovation – A New Approach to Government Innovation Policy.** The Industrial Institute for Economic and Social Research, Stockoholm.

Gittleman, M. – Wolff, E. (1995) **R&D Activity and Cross-Country Growth Comparisons.** Cambridge Journal of Economics 19, 189–207.

Glick, R. (1982) **R&D Effort and US Exports and Foreign Affiliate Production of Manufactures.** Research Policy 11, 359–372.

Gowdy, J.M. (1993) **Innovation Spending and Productivity Growth in the German Economy 1980–1986.** Applied Economics 25, 675–680.

Griliches, Z. (1990) **Patent Statistics as Economic Indicators: A Survey.** Journal of Economic Literature 28, 1661–1707.

Grossman, G.M. – Helpman, E. (1990) **The "New" Growth Theory – Trade, Innovation and Growth.** American Economic Reviw 80, 86–91.

Grossman, G.M. – Helpman, E. (1991a) **Trade, Knowledge, Spillovers and Growth.** European Economic Review 35, 517–526.

Grossman, G.M. – Helpman, E. (1991b) **Innovation and Growth in the Global Economy.** The MIT Press.

Guellec, D. – Ralle, P. (1993) **Innovation, Propriété Intellectuelle, Croissance.** Revue économique 44, 319–334.

Hall, B.H. – Griliches, Z. – Hausman, J.A. (1986) **Patents and R&D – Is There a Lag?** International Economic Review 27.

Hannan, T.H. – McDowell, J.M. (1984) **Market Concentration and the Diffusion of New Technology in the Banking Industry.** The Review of Economics and Statistics, Vol LXVI, 686–691.

Harris, C. – Vickers, J. (1985) **Perfect Equilibrium in a Model of a Race.** Review of Economic Studies 52, 193–209.

Hausman, J.A. – MacKieMason, J.K. (1992) **Price Discrimination and Patent Policy.** Rand Journal of Economics 19, 253–265.

Heertje, A. (1977) **Economics & Technical Change.** Weidenfeld & Nicolson, London, (First published in Dutch in 1973).

Hicks, J.R. (1963) **The Theory of Wages.** London, First published in 1932.

Hirsch, S. – Bijaoui, I. (1985) **R&D Intensity and Export Performance: A Micro View.** Weltwirtschaftliches Archiv 121, 238–251.

Hjerppe, R. (1985) **Suomen talous 1860–1985, kasvu ja rakennemuutos.** (The Finnish Economy in 1860–1985, growth and structural change), Suomen Pankki.

Horowitz, A.W. – Lai, E.L-C. (1996) **Patent Length and the Rate of Innovation.** International Economic Review, Vol 37, 785–.

Horstmann, I. – MacDonald, G.M. – Slivinski, A. (1985) **Patents as Information Transfer Mechanisms: To Patent or (Maybe) Not to Patent.** Journal of Political Economy 93, 837–858.

Humphrey, D.B. (1985) **Cost and Scale Economies in Bank Intermediation.** In Handbook of Banking Strategy (Ed. By Aspinvall & Eisenbeis), New York, Wiley & Sons.

Ito, K. – Puick, V. (1993) **R&D Spending, Domestic Competition, and Export Performance of Japanese Manufacturing Firms.** Strategic Management Journal 14, 61–75.

Johansen, S. – Juselius, K. (1990) **Maximum Likelihood Estimation and Inference on Cointegration – with Applications to the Demand for Money.** Oxford Bulletin of Economics and Statistics 52, 2/90, 169–210.

Jorgenson, D.W. – Gollop, F.M. – Fraumeni, F.M. (1987) **Productivity and U.S. Economic Growth.** Cambridge.

Jutikkala, E. (1953) **Uudenajan taloushistoria.** (The Economic History of the New Era), Turku.

Kaila, E. (1906) **Šekki.** Article in Tietosanakirja Encyclopedia, Helsinki.

Kalliala, K.J. (1958) **Säästöpankkien Keskus-Osake-Pankki 1908–1958.** Helsinki.

Kamien, M. – Schwartz, N. (1975) **Market Structure and Innovation – A Survey.** Journal of Economic Literature 13, 395–410.

Karafoulas, S. – Mantakas, G. (1996) **A Note on Cost Structure and Economies of Scale in Greek Banking.** Journal of Banking & Finance 20, 377–387.

Katz, M.L. – Shapiro, C. (1985) **Network Externalities – Competition and Compatibility.** American Economic Review.

Katz, M.L. – Shapiro, C. (1986) **Product Compatibility in a Market with Technological Progress.** Oxford Economic Papers, Vol 38, supplement, 146–165.

Kirzner, I.M. (1985) **Discovery and the Capitalist Process.** The University of Chicago Press.

Klette, T. – Meza, D. de (1986) **Is the Market Biased Against Risky R&D?** RAND Journal of Economics 17, 133–139.

Koponen, R. – Soramäki, K. (1998) **Intraday Liquidity Needs in a Modern Interbank Payment System – A Simulation Approach.** Bank of Finland Studies E:14.

Korpisaari, P. (1920) **Suomen pankit – Niiden kehitys, rakenne ja toiminta-muodot.** (Finnish Banks – their development, structure and activities) Kansantaloudellinen yhdistys, Helsinki.

Korpisaari, P. (1930) **Raha ja Pankit.** (Money and Banks) WSOY, Porvoo.

Kotabe, M.A. (1992) **Comparative Study of U.S. and Japanese Patent Systems.** Journal of International Business Studies, January-March, 147–168.

Kuusterä, A. (1995) **Säästöpankit suomalaisessa yhteiskunnassa 1822–1994.** (Savings Banks in the Finnish Society 1822–1994), Otava, Helsinki.

Laffont, J.-J. – Rey, P. – Tirole, J. (1997) **Competition between Telecommunications Operators.** European Economic Review 41, 701–711.

Lang, G. – Welzel, P. (1996) **Efficiency and Technical Progress in Banking – Empirical Results for a Panel of German Cooperative Banks.** Journal of Banking & Finance 20, 1003–1023.

Lanjouw, J.O. (1998) **Patent Protection in the Shadow of Infringement: Simulation Estimations of Patent Value.** Review of Economic Studies 65, 671–719.

Leppälahti, A. – Åkerblom, M. (1991) **Industrial Innovation in Finland – An Empirical Study.** Central Statistical Office of Finland, Studies 184, Helsinki.

Levin, R.C. – Klevorick, A.K. – Nelson, R.C. – Winter, S.G. (1987) **Appropriating the Returns from Individual Research and Development.** Brookings Papers on Economic Activity 3, 783–820.

Liebowitz, S.J. – Margolis, S.E. (1994) **Network Externality: An Uncommon Tragedy.** Journal of Economic Perspectives, Vol. 8, No 2, 133–150.

Lippman, S.A. – McCardle, K. (1988) **Preemption in R&D Races.** European Economic Review 32, 1661–1669.

Loury, G. (1979) **Market Structure and Innovation.** Quarterly Journal of Economics XCIII, 395–410.

Lovio, R. (1985) **Patentit ja korkean teknologian kauppa teknologia-indikaattoreina.** (Patents and high-tech trade as technology indicators), Technical Research Centre of Finland – Research Notes 408.

MacKinnon, J.G. (1991) **Critical Values for Cointegration Tests.** In Engle & Granger (eds): Long-run economic relationships – readings in cointegration, Oxfors.

Madarász, A. (1991) **Schumpeter's Theory of Economic Development.** In J.A. Schumpeter – Critical Assessments, 218–240, London. (Originally published: Acta Oeconomica 25 (1980), 337–356.)

Mansfield, E. (1969) **Industrial Research and Technological Innovation – An Econometric Analysis.** London.

Mansfield, E. – Schwartz, M. – Wagner, S. (1981) **Imitation Costs and Patents: An Empirical Study.** The Economic Journal 91, 907–918.

Mansfield, E. (1986) **Patents and Innovation – An Empirical Study.** Management Science 32, 173–181.

Mansfield, E. (1988) **Industrial R&D in Japan and the United States: A Comparative Study.** American Economic Review 78, 223–228.

Marjit, S. – Beladi, H. (1998) **Product Versus Process Patents: A Theoretical Approach.** Journal of Policy Modeling 20, 2, 193–199.

Matutes, C. – Padilla, J.A. (1994) **Shared ATM Networks and Banking Competition.** European Economic Review 38, 1113–1138.

Maudos, J. (1995) **Technical Change, Costs and Scale Economies in the Spanish Savings Banks.** The Automatic Teller Machine; Institute of European Finance, Research Paper 95/13.

McAndrews, J. – Roberds, W. (1997) **A Model of Check Exchange.** Federal Reserve Bank of Philadelphia Working Paper 97-16.

McKillop, D. – Glass, J.C. – Morikawa, Y. (1996) **The Composite Cost Function and Efficiency in Giant Japanese Banks.** Journal of Banking & Finance 20, 1651–1671.

Metcalfe, J.S. (1994) **Evolutionary Economics and Technology Policy.** The Economic Journal 104, 931–944.

Muto, S. (1987) **Possibility of Relicensing and Patent Protection** European Economic Review 31, 927–945.

Narin, F. – Noma, E. – Perry, R. (1987) **Patents as Indicators of Corporate Technological Strength.** Research Policy 16, 143–155.

Nelson, R.R. (1959) **The Simple Economics of Basic Scientific Research.** Journal of Political Economy, 297–306.

Noulas, A. – Miller, S.M. – Ray, S.C. (1993) **Regularity Conditions and Scope Estimates: The Case of Large-Sized U.S. Banks.** Journal of Financial Services Research, 235–248.

Oberender, P. von – Fricke, F.-U. (1993) **Möglichkeiten und Grenzen einer Europäischen Forschungs- und Technologiepolitik: Eine Ordnungs-teoretische Analyse.** Ifo Studien 3–4:39, 327–348.

OECD (1998) **Science, Technology and Industry Outlook.**

Pakes, A. – Griliches, Z. (1986) **Patents and R&D at the Firm Level: A First Look.** In Griliches (1984) "R&D, Patents and Productivity", The University of Chicago Press.

Pavitt, K. (1982) **R&D, Patenting and Innovative Activities.** Research Policy 11, 33–51.

Pennings, J. – Harianto, F. (1992) **The Diffusion of Technological Innovation in the Commercial Banking Industy.** Strategic Management Journal 13, 29–46.

Perelman, S. (1995) R&D, **Technological Progress and Efficiency Change in Industrial Activities.** Review of Income and Wealth 41, Nr 3, 349–366.

Pulley, L. – Braunstein, Y. (1992) **A Composite Cost Function for Multiproduct Firms with an Application to Economies of Scope in Banking.** The Review of Economics and Statistics 74, 29–46.

Reingaum, J. (1982) **A Dynamic Game of R&D – Patent Protection and Competitive Behavior.** Econometrica 50, 671–688.

Rivera-Batiz, L.A. – Romer, P.M. (1991a) **International Trade with Endogenous Technological Change.** European Economic Review 35, 971–1004.

Rivera-Batiz, L.A. – Romer, P.M. (1991b) **Economic Integration and Endogenous Growth.** Quarterly Journal of Economics 106, 531–555.

Romano, R.E. (1989) **Aspects of R&D Subsidization.** The Quarterly Journal of Economics, November, 863–873.

Romano, R.E. (1991) **The Optimal R&D Policy – Patents, Public Funding or Both?** Southern Economic Journal 57, 703–718.

Romer, P. (1989) **Endogenous Technological Change.** NBER Working paper 3210.

Rosegger, G. (1988) **From Hammer to Electric Generator: Technological Advance in Schmoller's Grundriß.** Journal of Institutional and Theoretical Economics 144, 591–600.

Rothwell, R. – Zegveld, W. (1981) **Industrial Innovation and Public Policy.** Frances Printer.

Saarenheimo, T. (1994) **Market Structure and the Propensity to Patent: Patenting or Secrecy.** In Studies on Market Structure and Tecnological Innovation, Bank of Finland B:49, Helsinki.

Saint-Paul, G. (1993) **Productivity Growth and the Structure of the Business Cycle.** European Economic Review 37, 861–890.

Schlegelmilch, B.B. (1988) **Der Zusammenhang Zwischen Innovationsneigung und Exportleistung.** Scmalenbachs Zeitschrift für Betriebswirtschaftliche Forschung, 40, 227–242.

Schmookler, J. (1966) **Invention and Economic Growth.** Harvard University Press.

Schumpeter, J.A. (1912) **Theorie der Wirtschaftlichen Entwicklung.** Leipzig.

Schumpeter, J.A. (1939) **Business Cycles – A Theoretical, Historical Statistical Analysis of the Capitalist Process.** New York.

Schumpeter, J.A. (1942) **Capitalism, Socialism and Democracy.** London.

Schwartz, M. (1991) **Patent Protection through Discriminatory Exclusion of Imports.** Review of Industrial Organization 6, 231–246.

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

Spence, M. (1984) **Cost Reduction, Competition and Industry Performance.** Econometrica 52, 101–121.

Stadler, G.W. (1990) **Business Cycle Models with Endogenous Technology.** American Economic Review, September 80, 763–778.

Stoneman, P. (1987) **The Economic Analysis of Technology Policy.** Clarendon Press.

Stoneman, P. – Diederen, P. (1994) **Technology Diffusion and Public Policy.** The Economic Journal 104, 918–930.

Takalo, T. (1996) **Innovation and Imitation under Imperfect Patent Protection.** University of Helsinki, Department of Economics, Discussion paper.

Tarkka, J. (1995) **Tax and Interest and the Pricing of Personal Demand Deposits.** In Approaches to Deposit Pricing, A Study of Deposit Interest and Bank Service Charges; Bank of Finland Studies E:2, Helsinki.

Taylor, M.S. (1994) **TRIPS, Trade and Growth.** International Economic Review 2:35, 361–381.

Trajtenberg, M. (1990) **A Penny for your Quotes: Patent Citations and the Value of Innovations.** Rand Journal of Economics 21, 172–187.

Urbans, R. (1963) **Suomen Säästöpankkilaitos 1822–1922.** (Finnish Savings Banks 1822–1922) Originally published in Swedish, Finnish translation by. A.V. Tola.

Usher, D. (1964) **The Welfare Economics of Invention.** Econometrica 31, 279–287.

Vesala, J. (1998) **Technological Transformation and Nonbank Competition in a Model of Retail Banking Duopoly.** Bank of Finland Discussion Papers 8/98, Helsinki.

Wright, B. (1983) **The Economics of Invention Incentives: Patents, Prizes and Research Contracts.** American Economic Review 73, 691–707.

Young, A. (1991) **Invention and Bounded Learning by Doing.** NBER Working Paper 3712.

Zif, J. – McCarthy, D. – Israeli, A. (1990) **Characteristics of Business with High R&D Investment.** Research Policy 19, 435–445.

Zimmermann, K.F. (1987) **Trade and Dynamic Efficiency.** Kyklos 40, 73–87.

Zimmermann, K.F. – Schwalbach, J. (1991) **Determinanten der Patentaktivität.** IFO-Studien – Zeitschrift für Empirische Wirtschaftsforschung 3–4:37.

Bank of Finland Studies E:18 • 2000

ISBN 951-686-651-4
ISSN 1238-1691

Gummerus Kirjapaino Oy
Jyväskylä 2000