Paris, June 19-20, 2008

Proceedings of the 2008 TELECOM ParisTech Conference on the Economics of Information and Communication Technologies (ICT)

TELECOM ParisTech - Paris, France - June 19-20, 2008

Editor : **Patrick Waelbroeck (TELECOM ParisTech)**
Organized by : **Department of Economics and Social Sciences of TELECOM ParisTech**

2008 : TELECOM ParisTech Conference ICT

ENST 2008 S 003

ENST 2008 S 003

TELECOM ParisTech - École Nationale Supérieure des Télécommunications

TELECOM ParisTech
Institut TELECOM - membre de ParisTech
46, rue Barrault - 75634 Paris Cedex 13
Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.fr
Département SES

Program of the 2008 Telecom ParisTech conference on the economics of ICT

Paris, June 19-20, 2008

Schedule

	Thursday
9:00-9:30	Welcome speech
9:30-10:30	Marshall Van Alstyne
11:00-12:30	MOBILE/IP1
12:30-14:00	Lunch
14:00-15:00	Stan Liebowitz
15:15-16:45	BANKING/TELECOM

Thursday evening, 7pm: conference dinner at Altitude 95 on first floor of the Eiffel Tower

	Friday
9:00-10:00	Bruno Jullien
10:30-12:00	PRODUCTIVITY/2-SIDED
12:00-13h30	Lunch
13:30-15:00	Iain Cockburn
15:15-16:45	BUNDLING/ IP2

Invited Talks

Marshall Van Alstyne (Boston University and MIT)
Discussant: Patrick Waelbroeck (Télécom ParisTech)

Stan Liebowitz (University of Texas, Dallas)
Discussant: Patrick Waelbroeck (Télécom ParisTech)

Bruno Jullien (Université de Toulouse, IDEI)
Discussant: Marc Bourreau (Télécom ParisTech)

Iain Cockburn (Boston University)
Discussant: Ron Zink (Microsoft), Dominique Guellec (OECD)

Sponsors

france telecom

INSTITUT TELECOM

Microsoft

Mobile

Pedro Pereira (Autoridade da Concorrencia) and Tiago Ribeiro (Indero), "A Model of Mobile Telephony with Policy Applications"
Discussant: Lionel Janin (DGTPE)

Øystein Foros (Norwegian School of Economics and Business Administration), Kåre P. Hagen (Norwegian School of Economics and Business Administration) and Hans Jarle Kind (Norwegian School of Economics and Business Administration), "Price-dependent Profit-Sharing as a Channel Coordination Device"
Discussant: Claire Chambolle (INRA and Ecole Polytechnique)

Intellectual Property 1

I.P.L. Png and Qiu-hong Wang (National University of Singapore), "Copyright Law, Movie Production, and Video Pricing: the European Rental Directive"
Discussant: Gilles Le Blanc (CERNA, Mines ParisTech)

Emeric Henry (London Business School) and Carlos J. Ponce (Universidad Carlos III de Madrid), "Waiting to Copy: the Dynamics of the Market for Technology"
Discussant: Matthieu Glachant (CERNA, Mines ParisTech)

Banking

Nicolas Serrano-Velarde (European University Institute), "The Financing Structure of Corporate R&D - Evidence from Regression Discontinuity Design"
Discussant: Frédérique Savignac (Banque de France)

Jason Allen (Bank of Canada), Robert Clark (HEC Montréal) and Jean-François Houde (University of Wisconsin-Madison), "Market Structure and the Diffusion of Electronic Banking"
Discussant: Marianne Verdier (Telecom ParisTech)

Ricardo Ribeiro (London School of Economics and Political Science), "An Empirical Model of Multi-venue Trading Competition post Mi-Fi"
Discussant: Lapo Filistrucchi (Tilburg University and University of Siena)

Telecom industries

Hans Friederiszick, Michal Grajek and Lars-Hendrik Röller (ESMT), "Analyzing the Relationship between Regulation and Investment in the Telecom Sector"
Discussant: Romain Lestage (Orange Labs)

Duarte Brito (Universidade Nova de Lisboa), Pedro Pereira (Autoridade da Concorrência) and João Vareda (Autoridade da Concorrência), "On the Regulation of Next Generation Networks"
Discussant: Matthieu Manant (Laboratoire Econometrie, CNAM)

Alessandro Avenali, Giorgio Matteucci, Pierfrancesco Reverberi (Sapienza, Università di Roma), "Vertical Separation and Network Investment in Telecommunications"
Discussant: Stefan Behringer (Universitat Frankfurt)

Productivity

María Rosalía Vicente Cuervo (University of Oviedo) and Maria do Rosário O. Martins (Universidade Nova de Lisboa, ISEG), "Information Technology, Efficiency and Productivity in SMEs: Evidence for Portugal"
Discussant: Thomas Houy (Telecom ParisTech)

Gaaitzen J. de Vries and Michael Koetter (University of Groningen), "How does ICT enhance Productivity? Evidence from Latent Retail Technologies in Chile"
Discussant: Maria do Rosario O. Martins (U. Nova Lisbao, ISEG)

Two-sided markets

Pierre Gazé and Anne-Gael Vaubourg (Université d'Orléans, LEO), "Electronic Intermediation and Two-sided Markets: what happens when Sellers and Buyers can switch?"
Discussant: Benoît Crutzen (Erasmus Universiteit Rotterdam)

Marc Bourreau and Marianne Verdier (Telecom ParisTech), "Private Cards and the Bypass of Payment Systems by Merchants"
Discussant: Benoît Crutzen (Erasmus Universiteit Rotterdam)

Bundling

Edmond Baranes and Jean-Christophe Poudou (University Montpellier 1, LASER), "Cost-Based Access Pricing and Collusion with Bundling"
Discussant: Antonin Arlandis (Orange Labs)

Grazia Cecere (Université de Paris Sud XI ADIS), "Triplay vs Software Voice in France"
Discussant: Joeffrey Drouard (Telecom ParisTech)

IP2

Jing-Yuan Chiou (IMT Lucca), "The Patent Quality Control Process: Can We Afford An (Rationally) Ignorant Patent Office"
Discussant: Serge Pajak (Telecom ParisTech)

Eric Darmon (Université Rennes 1), Alexandra Rufini and Dominique Torre (University of Nice Sophia-Antipolis), "Back to Software Profitable Piracy: The Role of Delayed Adoption and Information Diffusion"
Discussant: Patrick Waelbroeck (Telecom ParisTech)

Contents

A Model of Mobile Telephony with Policy Applications*

(PRELIMINARY VERSION, DO NOT QUOTE OR CITE)

Pedro Pereira† Tiago Ribeiro‡

Autoridade da Concorrência Indera

December 2007

Abstract

In this article, we develop a model of the mobile telephony industry, that includes both a demand and a supply side. The model is estimated for a rich panel of firm level Portuguese data, and used to perform several policy exercises. We simulate the effect of the merger that would reduce the number of firms from three to two on prices and social welfare. Our results indicate that the merger would lead to substantial price increases. On average, each household would spend an additional 6.3% of the current expenditure levels. The comparison of observed and estimated margins suggests that the Nash assumption is plausible. The merger seems to generate small efficiency gains. Marginal cost reductions of 10% would generate small price reductions. The entry of a firm after the merger would lead to a less competitive equilibrium, than before the merger. The entry of a fourth firm (...).

Key Words: *Mobile Telephony, Merger, Prices, Efficiency Gains, Entry*

JEL Classification: L13, L43, L93

*We thank D. Brito for useful comments. The opinions expressed in this article reflect only the authors' views, and in no way bind the institutions to which they are affiliated.

†AdC, Rua Laura Alves, n^o 4, 6^o, 1050-188 Lisboa, Portugal, e-mail: jppereira@oninetspeed.pt

‡Indera - Estudos Económicos, Lda, Edifício Península, Praça Bom Sucesso, 127/131, Sala 202, 4150-146 Porto, Portugal, e-mail: tiago.ribeiro@indera.pt

1 Introduction

In this article, we develop a structural model of the mobile telephony industry, that includes both the demand and the supply side. The model is estimated for a rich panel of firm level Portuguese data, and used to perform several policy exercises. In Portugal there are three mobile network operators, *Tmn*, *Vodafone*, and *Optimus*, which in 2005 had revenue market shares of 50%, 37% and 13%, respectively.

Our consumer decision model has two components: **(i)** the sampling process, and **(ii)** the consumer structural decision model. For the sampling process, we assume that entry into the market by consumers follows an S-shaped diffusion process.[1] For the consumer structural decision model, we assume a discrete choice model. For the cost model, we assume a quadratic cost function.

We take advantage of the richness of our data set in the specification of the demand and cost models. The demand model includes both mobile and fixed telephony products. For mobile telephony we consider two products: a pre-paid cards product, and a contract product. For fixed telephony we consider also two products: the product of the incumbent *PT Comunicações*, *PTC*, and an aggregate product for the entrants in fixed telephony. In addition, we include the prices of *SMS* as a characteristic of the products of mobile telephony firms. In the cost functions, we consider the prices of four production factors: labor, capital, materials, and interconnection.

We use the demand model to estimate the price elasticities of demand. The demand model on which we base our conclusions is a nested logit model. Consumers are quite sensitive to price variations in mobile telephony. We use the cost model to estimate: the marginal costs, the average costs, and the economies of scale.

We use the structural model to perform three policy exercises. In the first policy exercise, we simulate the effect of a merger between *Tmn* and *Optimus*. Assuming firms play a Bertrand game, we use the demand elasticities to estimate the marginal costs. The comparison of observed and estimated margins suggests that the assumption of Nash behavior is plausible. Given the demand and cost estimates, we simulate the effect of a merger on prices and social welfare. Our results indicate that the merger would lead to substantial price increases. On average, each consumer would spend an additional 6.3% on mobile communications, compared to current expenditure levels. The clients of *Optimus* would be disproportionately affected.[2] On average, the consumer surplus per minute would decrease

[1] Alternatively, one could assume that the evolution of the characteristics of mobile telephony with respect to fixed telephony explains fully the evolution of the market shares, and take to the data a simple discrete choice model. Although in the present case both alternatives would yield very similar results, they are, nevertheless, conceptually very different.

[2] Given the estimated elasticities, the merged firm would equate the prices of similar products to levels

by $6.332 \cdot 10^{-3}$ euros, and the profits per minute would increase by $6.642 \cdot 10^{-3}$ euros. The cost estimates suggest that the merger would generate small efficiency gains, if any. Since the [3] estimated marginal costs are low compared to the prices, potential marginal cost decreases of up to 10% would have little impact on prices.[3]

In the second policy exercise, we simulate the effect of the entry of a firm with the characteristics of *Optimus* after the merger of *Tmn* and *Optimus*. If entrant sets prices equal to the post-merger industry average and the rivals do not react, it attains a market share of at most 5%. In a new equilibria, the market share of the entrant would be around 10%.

Third, we simulate the effect of the entry of a firm with the characteristics of *Optimus* without the merger of *Tmn* and *Optimus*.

Our methodological approach draws on the discrete choice literature, represented among others by Domencich and McFadden (1975), Mcfadden (1974), McFadden (1978), and McFadden (1981). In the industrial organization literature, Berry (1994), Berry, Levinsohn, and Pakes (1995), and Nevo (2001) applied discrete choice models to the analysis of market structure. Dube (2005), Ivaldi (2005), Ivaldi and Verboven (2005), Nevo (2000), and Pinkse and Slade (2004) analyzed the impact of a merger in a framework similar to ours.[4] These studies used aggregate data, with the exception of Dube (2005), which used household level data.

Regarding the empirical literature on mobile telephony, Parker and Roeller (1997) use US data from 1984 to 1988 to estimate a structural model of the mobile telephony industry. They report an own-price elasticity of demand of -2.5, and increasing marginal costs. Using the same data, Miravete and Roeller (2004) estimate an equilibrium model of horizontal product differentiation where firms compete in nonlinear tariffs. They report constant marginal costs. Madden and Dalzell (2004) use annual panel data for 56 countries from $1995 - 2000$. They estimate an own-price elasticity of -0.55 and an income elasticity of 4.76. They also estimate network effects. Hausman (1997) reports an own-price elasticity of subscription of -0.51 for cellular subscription in the 30 largest US markets over the period $1988 - 1993$. Hausman (2000) using more recent data reports an own-price elasticity of subscription of -0.71. Gagnepain and Pereira (forthcoming) studied the effect of entry of *Optimus* in 1997 on costs and competition in the Portuguese mobile telephony industry. The results suggested

closer to today's prices of *Tmn*.

[3] We considered the case where the merger could generate efficiency gains. However, if the firms in the industry face moral hazard problems, such as those analyzed by Gagnepain and Pereira (forthcoming), the decrease in competitive pressure caused by the merger could lead firms to lower their cost reducing efforts, and thereby lead to higher marginal costs. See also Brito and Pereira (2007).

[4] See also Baker and Bresnahan (1985) and J. Hausman and Zona (1994).

that the entry of a third operator in 1998 lead to significant cost reductions and fostered competition. The authors construct and estimate a model that includes demand, network, and cost equations. The latter accounts for inefficiency and cost reducing effort. Grzybowski and Pereira (2007) used a simple aggregate nested logit model with network effects, with market shares in terms of subscribers. Their results indicate that the merger would lead to price increases of $7 - 10\%$. Okada and Hatta (1999) using annual Japanese data from 1992 to 1996, totaling 235 observations, estimated an almost ideal demand system. They report an own-price elasticity of demand for mobile telephony of -3.963 and -1.405, respectively, a cross-price elasticity of the demand of mobile telephony with respect to the price of fixed telephony of 0.866, and a cross-price elasticity of the demand for fixed telephony with respect to the price of mobile telephony of 0.276. Rodini and Woroch (2003) use a US household annual survey for the period 2000 to 2001, with 327.920 observations to estimate own and cross price elasticities of mobile and fixed telephony. Estimated cross-price elasticities show that a second line and mobile services are substitutes of one another. They estimate an own-price elasticity of mobile access demand of -0.43, an own-price elasticity of mobile access and usage of -0.60 and a cross-price elasticity of mobile demand with respect to fixed access of 0.13.

The rest of the article is organized as follows. Section 2 gives an overview of the Portuguese mobile telephony industry. Section 3 describes the data. Section 4 presents the model. Section 5 describes the econometric implementation, and presents the basic estimation results. Section 6 analyzes the impact of the merger, and Section 7 concludes.

2 Overview of the Portuguese Industry

In Portugal, the firm associated with the telecommunications incumbent, *Tmn*, started its activity in 1989 with the analogue technology *C-450*. In 1991, the sectorial regulator, *ICP-ANACOM*, assigned two licenses to operate the digital technology *GSM 900*. One of the licenses was assigned to *Tmn*. The other license was assigned to the entrant *Vodafone*. *Tmn* introduced pre-paid cards in 1995 for the first time worldwide. In 1997, the regulator assigned three licenses to operate the digital technology *GSM 1800*. Two licenses were assigned to *Tmn* and *Vodafone*. A third license was assigned to the entrant *Optimus*, which was also granted a license to operate *GSM 900*. In 2001, *ICP-ANACOM* assigned licences to operate the *3G* technology *IMT2000/UMTS*. Three licenses were assigned to *Tmn*, *Vodafone*, and *Optimus*. A fourth license was assigned to the entrant *Oniway*, which

was not granted a license to operate *GSM*, and never operated. Service began in 2003.[5]

[*Figure* 1]

After its inception in 1989, the Portuguese mobile telephony industry had a fast diffusion, analyzed in Gagnepain and Pereira (2007) () and Pereira and Pernias (2006). In 2005 the penetration rate of mobile telephony in Portugal was 110%. After entering the market in 1992, *Vodafone* gained revenue market share rapidly. During the duopoly period, i.e., from 1992 to 1997, *Tmn* and *Vodafone* essentially shared the market. The entry of *Optimus* led to an asymmetric split of the market, which suggests that this event had a significant impact in the industry, illustrated in Figure 1. A similar perspective can be gleaned from the analysis of the time series of average prices of *Tmn* and *Vodafone*, presented in Figure 1. The average prices of *Tmn* and *Vodafone* move in parallel, and have a downward break in 1997. This suggests that the entry of *Optimus* in 1998 caused the rivals to reduce prices.[6]

In February 2006, the holding company *Sonaecom*, which owns *Optimus*, made a hostile take-over bid for the holding company *Portugal Telecom*, the telecommunications incumbent, which owns *Tmn*. The transaction required the approval of the *Portuguese Competition Authority*. *Sonaecom* justified the merger of *Tmn* and *Optimus* on the basis of: substantial putative efficiency gains, and the inability of the firms increasing prices under the current market conditions. The *Portuguese Competition Authority* approved the transaction with six remedies in mobile telephony. First, the merged firm would return to *ANACOM* the licenses to use the *GSM* and the *UMTS* spectrum of either *Tmn* or *Optimus*. Second, the merged firm would develop a wholesale reference offer for mobile virtual network operators. Third, there would be a financial compensation scheme, intended to overcome the price mediated network externalities faced by an entrant mobile network operator. Fourth, the merged firm would limit the differences between the on-net and off-net prices with respect to any entrant, mobile network operator or mobile virtual network operator. Fifth, the merged firm would take steps to reduce the customer switching costs in mobile telecommunications. Sixth, the merged firm would be subject to a price-cap. However, the transaction did not go through because the shareholders of *PT* voted against changing a clause of the statutes of the firm limiting the voting rights of the shareholders, a prerequisite for the operation.[7]

[5]All of the licenses for *GSM* 900 and for *GSM* 1800 were assigned through public tenders, following EU Directives 91/287 and 96/2, respectively. The first Directive instructed Member States to adopt the *GSM* standard, and the second to grant at least 2 *GSM* 900 licenses and to allow additional firms to use GSM 1800. System *GSM* 900 operates on the 900 MHz frequency. System *GSM* 1800 operates on the 1800 MHz frequency. The licenses for *3G* were assigned through public tenders, following EU Decision 128/1999/EC.

[6]Note that economic theory is not always conclusive regarding the relation between the number of competitors in a specific industry and firms' prices. Garcia et al. (2006), Rosenthal (1980), and Seade (1980) develop models where prices increase with the number of firms in the market.

[7]The statutes of *PT* imposed that no shareholder could have more than 10% of the voting rights,

3 Econometric Model

In this section, we present the econometric model. First, we provide a brief introduction of the demand and cost models we estimate. Second, we describe the implications of these models for the welfare analysis. Third, we present the assumptions about the behavior of firms.

3.1 Demand

3.1.1 Utility of Telephony Services

A consumer chooses among a set of alternative products for mobile and fixed telephony. The products differ in: **(i)** the price, **(ii)** the type of subscription of mobile telephony, i.e., pre-paid card or contract, **(iii)** the size of the network of the firm, and **(iv)** the price of *SMS* of the firm. We assume that the size of the network and price of *SMS* are not relevant for fixed telephony, and set these values to zero in fixed telephony products.

We omit subscripts whenever possible. In period $t = 1, ..., T$ consumers derive from alternative $i = 1, ..., I$ utility:

$$U_i(p_i, x_i, \theta) = V_i(p_i, x_i, \theta) + \varepsilon_i,$$

where p_i is the price of alternative i, x_i is a J dimensional vector of the other characteristics of alternative i, θ is a vector of parameters, and finally, ε_i is a random disturbance independent across consumers and time, and identically distributed. We assume additionally that:

$$V_{ni}(p_i, x_i, \theta) := p_i\alpha + g(x_i, \beta),$$

where

$$g(x_i, \beta) := \sum_{j=1}^{J} x_{ij}\beta_j,$$

$$\theta := (\alpha, \beta),$$

and where α is the price coefficient, i.e., the negative of the marginal utility of income. Expression $g(\cdot)$ is a linear combination that summarizes the utility component associated with all product characteristics other than price. The parameters β translate the consumer valuation of the different product characteristics. This formulation encompasses all the models analyzed in this paper. If ε_i has an extreme value Type I distribution, one obtains the standard multinomial logit model. Setting the joint distribution of ε_i to be of the generalized extreme value family, with the required generating function, one obtains the nested logit model.

irrespective of the number of shares owned.

3.1.2 Choice Probabilities

A consumer chooses product i if $U_i > U_j$, for all $j \neq i$. This occurs with probability: [7]

$$P_i := \Pr\left[V_i - V_j + \varepsilon_i > \varepsilon_j, \text{ for all } j \neq i, j = 1, \ldots, I\right] = \int F_i(V_i - V_1 + u, \ldots, u, \ldots, V_i - V_I + u)du,$$

where $F(\cdot)$ is the joint distribution function of $(\varepsilon_1, \ldots, \varepsilon_I)$, and $F_i(\cdot)$ is its partial derivative with respect to the i^{th} argument.

If $F(\cdot)$ is an extreme value type I distribution, with the generating function $H\left(x_1, \ldots, x_J\right) = \sum_{j=1}^{J} x_j$, one obtains the standard multinomial logit expression for the choice probabilities:

$$P_i = \frac{e^{V_i}}{\sum_j e^{V_j}}.$$

If $F(\cdot)$ is a generalized extreme value joint distribution, with the generating function $H\left(x_1, \ldots, x_J\right) = \sum_{k=1}^{K} \left(\sum_{j \in B_k} x_j^{\frac{1}{\lambda_k}}\right)^{\lambda_k}$, one obtains the nested logit model:

$$P_i = \frac{e^{\frac{V_i}{\lambda_k}}}{\sum_{j \in B_k} e^{\frac{V_j}{\lambda_k}}} \frac{\left(\sum_{j \in B_k} e^{\frac{V_j}{\lambda_k}}\right)^{\lambda_k}}{\sum_l \left(\sum_{j \in B_l} e^{\frac{V_j}{\lambda_l}}\right)^{\lambda_l}}, \quad i \text{ on nest } k.$$

3.1.3 Aggregate Market Shares

With aggregate data, it is common to express market shares as a linear function of the indirect utilities. Let product i belong to nest k, and product 1 belong to nest 1. Denote by $P_{i|k}$, the choice probability, given that we restrict choices to the products in a nest. The so-called inversion of market shares is given by expressions (1) and (2) for the multinomial logit and nested logit models, respectively:

$$\log\left(\frac{P_i}{P_1}\right) = V_i - V_1, \tag{1}$$

$$\log\left(\frac{P_i}{P_1}\right) = V_i - V_1 + (1 - \lambda_k)\log(P_{i|k}) - (1 - \lambda_1)\log(P_{1|1}), \tag{2}$$

where

$$P_{i|k} := \frac{e^{V_i/\lambda_k}}{\sum_{j \in B_k} e^{V_j/\lambda_k}}.$$

If the baseline product, denoted by 1, is the only element in its nest, then λ_1 is not identified and is normalized to 1. This happens in some of the models we estimate.

3.1.4 Observed Market Shares

We do not observe the consumers' choices directly. The observed market shares are a result [8] of the choices of several consumers that in the past decided to buy mobile telephony services, and the choices of several consumers that have not yet decided to buy mobile telephony services. As these last consumers enter the market, the diffusion of mobile telephony unfolds.

According to this view, we model the observed demand as having two components: **(i)** a diffusion process describing the evolution of the market from the inception of mobile telephony, up to the equilibrium between mobile and fixed services, and **(ii)** a discrete choice model for the equilibrium market shares.

Denote by P^0, the vector of market shares before mobile telephony was introduced, where the first element, fixed line P_1^0, is 1, and all the others elements are 0, and denote by P^1, the vector of equilibrium market shares. By equilibrium market shares we mean the market shares that would result if everyone chose their preferred product without any switching costs. Denote by $D(t)$, the normalized diffusion curve; $D(t) = \frac{N(t)}{\kappa}$ in the notation of Pereira and Pernias (2006). The expression $D(t)$ is a reduced form of either the decision process described, or of the evolution due to network effects. Denote by $P(t)$, the observed market shares at time t. The expression $P(t)$ results form a fraction $D(t)$ of the population having chosen according to P^1, and a fraction $1 - D(t)$ not having made any decision yet. Therefore:

$$P(t) \simeq P^0(1 - D(t)) + P^1 D(t).$$

The expression of $P(\cdot)$ is an approximation because P^1 depends on variables that change over time, most notably the size of the network. This simplification is meant to express the assumption that most of the time the evolution of the market shares is driven by the diffusion of mobile telephony.

The observed share of product i is then:

$$P_i(t) = P_i^1 D(t);$$

and the observed share of product 1, the product of PTC is:

$$P_1(t) = 1 - D(t) + P_1^1 D(t).$$

Combining the previous expressions we obtain the ratio of observed market shares:

$$\frac{P_i(t)}{P_1(t)} = \frac{P_i^1 D(t)}{1 - D(t) + P_1^1 D(t)} = \frac{P_i^1}{P_1^1} \frac{1}{1 + \frac{D(t)}{1-D(t)} \frac{1}{P_1^1}}.$$

If P_1^1 does not change much over time, then the denominator is just a function of t. If we take $D(t)$ to be the normalized logistic diffusion curve, i.e.,

$$D(t) = \frac{1}{1 + \exp(\gamma_1 + \gamma_2 t)},$$

then we have:

$$\frac{D(t)}{1-D(t)}\frac{1}{P_1^1} = \exp(\gamma_1 + \gamma_2 t - \log(P_1^1)), \qquad 9$$

and

$$\begin{aligned} \log\left(\frac{P_i(t)}{P_1(t)}\right) &= \log\left(\frac{P_i^1}{P_1^1}\right) - \log(1+\exp(\tilde{\gamma}_1 + \gamma_2 t)), & (3)\\ &= \log\left(\frac{P_i^1}{P_1^1}\right) + h(t). & (4) \end{aligned}$$

The first term in (3) was derived in the previous section, and $h(\cdot)$ is an almost linear function in t. In this model, $h(\cdot)$ does not have an interpretation of an utility component. It is a correction term to account for the observational process.

3.1.5 Price Elasticities of Demand

Denote by ε_{ij}, the elasticity of demand of product i with respect to the price of product j:

$$\varepsilon_{ij} := \frac{\partial P_i}{\partial p_j}\frac{p_j}{P_i}.$$

In the multinomial logit model, the partial derivative is:

$$\frac{\partial P_i}{\partial p_j} = \begin{cases} \alpha P_i(1-P_i) & \text{if } i = j \\ -\alpha P_i P_j & \text{otherwise;} \end{cases}$$

implying the following elasticities:

$$\varepsilon_{ij} = \begin{cases} \alpha p_i(1-P_i) & \text{if } i = j \\ -\alpha p_j P_j & \text{otherwise.} \end{cases}$$

In the nested logit model the partial derivatives are:

$$\frac{\partial P_i}{\partial p_j} = \begin{cases} \alpha P_i\left[(1-\frac{1}{\lambda_k})P_{i|k} - P_i + \frac{1}{\lambda_k}\right] & \text{if } i=j; \quad i \text{ on nest } k \\ \alpha P_i\left[(1-\frac{1}{\lambda_k})P_{j|k} - P_j\right] & \text{if } i\neq j; \quad i,j \text{ on nest } k \\ -\alpha P_i P_j & \text{if } i\neq j; \quad i,j \text{ in different nests;} \end{cases}$$

implying the following elasticities:

$$\varepsilon_{ij} = \begin{cases} \alpha p_i\left[(1-\frac{1}{\lambda_k})P_{i|k} - P_i + \frac{1}{\lambda_k}\right] & \text{if } i=j; \quad i \text{ on nest } k \\ \alpha p_j\left[(1-\frac{1}{\lambda_k})P_{j|k} - P_j\right] & \text{if } i\neq j; \quad i,j \text{ on nest } k \\ -\alpha p_j P_j & \text{if } i\neq j; \quad i,j \text{ in different nests.} \end{cases}$$

3.1.6 Consumer Welfare Valuation

Denote by $V_j^{'}$ and $V_j^{''}$, the utility levels before and after the merger, respectively. The[10] merger implies three types of changes. First, prices change, which requires computing the market equilibrium after the merger. Second, the characteristics of the products change, i.e., x_i changes. Third, the number of products offered may change.

The generalized extreme value model, of which the multinomial and the nested logit models are particular cases, provides a convenient computational formula for the exact consumer surplus, up to a constant, associated with a policy that changes the attributes of the products in the market. This expression, known as the "log sum" formula, is:[8]

$$\Delta CS_n = \frac{1}{\alpha}\left[\ln H\left(e^{V_{n1}^{''}},\ldots,e^{V_{nJ}^{''}}\right) - \ln H\left(e^{V_{n1}^{'}},\ldots,e^{V_{nJ}^{'}}\right)\right]. \tag{5}$$

This formula is valid only when the indirect utility function is linear in income, i.e., when price changes have no income effects, which is the case assumed here.

3.2 Supply

3.2.1 Cost of Mobile Telephony

We index firms with subscript $i = tmn, vod, opt$, with the obvious interpretation. Denote by ω_{jit}, the price of production factor j for firm i in period t. Labor, capital, materials, and interconnection are indexed respectively by $j = l, k, m, a$. The cost function of firm i is:

$$\log(c_{it}) = \alpha_0 + \alpha_y \log(y_{it}) + \alpha_{yy}\log(y_{it})^2 + \sum_{j=l,k,m,a}\gamma_j \log(\omega_{jit}) + \delta t + \varepsilon_{it}. \tag{6}$$

Expression (6) is a simplified version of a translog cost function, where with the exception of $\log(y_{it})^2$, all cross terms were set to zero. From (6), we compute the *economies of scale*, defined as the ratio of marginal to average costs:

$$EcS_{it} := \alpha_y + 2\alpha_{yy}\log(y_{it});$$

and the marginal costs:

$$MgC_{it} := EcS_{it}\frac{c_{it}}{y_{it}}.$$

[8] This expression was developed by Domencich and McFadden (1975), and Mcfadden (1974) for the multinomial logit model, and by McFadden (1978) and McFadden (1981) for the nested logit model. Small and Rosen (1981) elaborate on the connection between the above measures of welfare and standard measures of consumer surplus.

3.2.2 Price Equilibrium

The profit function of firm i is:

$$\Pi_i = \sum_{j=1}^{J} \delta_{ij}\pi_j,$$

where $\pi_j := p_j Q_j(\mathbf{p}) - C_j(Q_j(\mathbf{p}))$ is the profit in market j, $\delta_{ij} = 1$ if firm i sells product j, and $\delta_{ij} = 0$ otherwise. We assume that firms choose prices and play a static non-cooperative game, i.e., a Bertrand game. The Nash equilibrium of the game is characterized by the following set of first order conditions:[9]

$$\sum_{i=1}^{I} \delta_{ik}\frac{\partial \Pi_i}{\partial p_k} = \sum_{i=1}^{I} \delta_{ik}\left[\delta_{ik}Q_k + \sum_{j=1}^{J}\delta_{ij}\frac{\partial Q_j}{\partial p_k}\left(p_j - \frac{\partial C_j}{\partial p_j}\right)\right] = Q_k + \sum_{j=1}^{J}\gamma_{kj}\frac{\partial Q_j}{\partial p_k}(p_j - c_j),$$

where $c_j := \frac{\partial C_j}{\partial p_j}$, and $\gamma_{kj} = 1$ if products j and k are sold by the same firm, and $\gamma_{kj} = 0$ otherwise.

Let matrices Γ and Φ consist of the elements $\Gamma_{ij} = \gamma_{ij}$ and $\Phi_{ij} = \frac{\partial Q_j}{\partial p_i}$, respectively. Matrix Γ represents the market structure, and matrix Φ consists of the demand estimates. Denote by $A \circ B$ the element by element product of matrices A and B, i.e., the Hadamard product. The system that defines the equilibrium can be written as:

$$\mathbf{Q} + (\Gamma \circ \Phi)(\mathbf{p} - \mathbf{c}) = 0. \tag{7}$$

Initially there are three mobile telephony firms: *Tmn*, *Vodafone*, and *Optimus*. Each firm controls two products: a pre-paid card product, and a subscription product. Thus:

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

In the matrix above, the first and the last products represented are fixed telephony from *PTC* and from other firms, respectively. The remaining products refer to mobile telephony.

In the course of the analysis, we will assume an alternative form for the matrix Γ, associated with the merger of *Tmn* and *Optimus*.

3.2.3 Profit Variation

Denote by $(\mathbf{Q}', \mathbf{p}')$ and $(\mathbf{Q}'', \mathbf{p}'')$ the levels of output and prices before and after the merger, respectively. Taking a first-order approximation of the cost function of product j

[9] We assume that a Nash equilibrium exists. Caplin and Nalebuff (1991) proved existence in a general discrete choice model, with single product firms. Anderson and de Palma (1992) proved existence for the nested logit model with symmetric multiproduct firms.

around the current level of output, the profit function is:

$$\pi_j(\mathbf{Q}, \mathbf{p}) = p_j Q_j - C_j(Q_j) \simeq p_j Q_j - C_j(Q'_j) - c_j(Q_j - Q'_j).$$

The profit variation for product j is then:

$$\begin{aligned} \Delta\pi_j = \pi_j(\mathbf{Q}'', \mathbf{p}'') - \pi_j(\mathbf{Q}', \mathbf{p}') &\simeq p''_j Q''_j - p'_j Q'_j - c_j(Q''_j - Q'_j) \\ &= (p''_j - c_j)Q''_j - (p'_j - c_j)Q'_j. \end{aligned}$$

4 Econometric Implementation

4.1 Data

The data consists of quarterly observations for the period 1992:1 − 2005:4. For the demand models, we use a panel for 2001:4 − 2005:4. For the models with products of the fixed telephony entrants the panel is unbalanced.

The variables were constructed as follows. In the cost function, total costs, c, production, y, wages, ω_l, price of materials, ω_m, access price, ω_a, and price of capital, ω_k, correspond to total costs in thousands of euros, originated voice traffic in thousands of minutes, total labor costs over number of employees, cost of supplies over originated voice traffic, termination costs over originated voice traffic, and interest rate of ten-year treasury bonds, respectively.

In the demand function, the price of product i on period t, p_{it}, is measured as total revenues over traffic supplied. Moreover, the size of the network of firm j, s_{jt}, is measured by the number of subscribers of firm j.

[*Figure* 2]

[*Figure* 3]

The raw data exhibits significant quarterly variation, which may reflect mostly accounting practices, and not the underlying evolution of the market. This is the case of the behavior of the average prices, which should evolve smoothly, and not exhibit quarterly variation of the magnitude present in the original data. In accordance with this interpretation, we removed the higher frequencies from our time series by means of kernel smoothing algorithms. In the series where it was appropriate, the data was first isotonized and afterwards smoothed. The series for originated minutes and subscribers where set to be in clear expansion. The comparison between raw and smoothed data is presented in Figures 2 and 3.

[*Figure* 4]

We classified the mobile telephony options into six products. For each of the three mobile telephony firms we consider: a pre-paid card product, and a contract product.[10] Consumers with pre-paid cards and contracts have different consumption patterns, as presented in Figure 4. We also classified the fixed telephony options into six products, each associated to one of the six largest fixed telephony firms: the incumbent *PTC*, *Cabovisão*, *Novis*, *Oni*, *Tele 2*, and *Vodafone*. In some models, the entrants in fixed telephony appear aggregated into one single option, as their individual relevance is small.

4.2 Cost Function Estimates

We estimated eight cost functions by OLS. The results are presented in Tables 1 and 2.

[*Table* 1]

[*Table* 2]

Models 1 to 4 include proxies for the prices of two productions factor: labor and capital. Models 1 and 2 contain only linear terms in the output variable. Models 3 and 4 have quadratic terms allowing for varying economies of scale. Models 2 and 4 allow for different time trends between firms, i.e., different technological progress across firms.

Models 5 to 8 reproduce models 1 to 4, and add proxies for the prices of two additional production factors: interconnection, and materials.[11]

The results are qualitatively the same across all models, with the exception of model 8, in which the average and marginal cost curvatures are reversed. However, the coefficients associated with the output variables are not statistically significant in this model. For each model, and for each firm, we computed: **(i)** the average costs, **(ii)** the marginal costs, and **(iii)** the economies of scale. The same quantities were also calculated for the hypothetical firm resulting from the merger of *Tmn* and *Optimus*, labeled *Tmn+Optimus*.

Across all models, *Tmn* has the lowest marginal and average costs, and benefits from higher economies of scale. *Optimus* and *Vodafone* have similar average and marginal costs. *Optimus* had a more substantial technological progress than the other two firms.

[*Figure* 5]

Where relevant, we computed the efficient scale, as well as the value of the marginal cost at this point. The average cost curve is very flat with respect to most of the observed

[10] With the exception of Miravete and Roeller (2004), the literature considers only one product per firm.

[11] We also computed several versions of a full translog model imposing in turn homogeneity and homotheticity conditions. These restriction were in general not supported by the data and the curvature with respect to output was reversed as in model 8. We adopted a simple model that approximated appropriatly the features of the cost function most relevant for this paper.

output.[12] As a consequence, the identification of the minimum of the average cost curve is very sensitive to the model specification, and points to extremely high and unlikely values of output. Several measures indicate that economies of scale are exhausted well before the computed minimum of average cost curve. We calculated the smallest quantity for which the hypothesis that marginal and average cost are equal cannot be rejected. These values, labeled with a superscript b, indicate that the efficient scale is much smaller that the point estimate. For model 7, economies of scale have already been exhausted by *Tmn*. Another measure that points in this direction is the value of the average cost at the minimum of the average cost curve, which is very close to *Tmn*'s average cost. Figure 5 illustrates these points. The horizontal lines refer to the average costs for *Optimus*, *Vodafone*, *Tmn*, *Tmn+Optimus*, and the estimated minimum, respectively, based on model 3. In the remainder of the article, we base our calculations on model 3.

Overall, the efficiency gains resulting from scale economies associated to the merger are likely to be small. Most of these gains would accrue to the products of *Optimus*, which would be produced at a lower marginal cost, benefiting from the scale of *Tmn*.

4.3 Demand Estimates

We estimated five models of the demand function expressed in equations (1) and (2), with the modification described in equation (4). The models where estimated by both by OLS and IV, to account for the possible endogeneity of prices. We used the following instruments: total costs, labor costs, materials costs, and interconnection costs. We describe these calculations in turn.[13] The results of demand estimates are presented in Tables 3, 4, and 5.

Table 3 presents models I and II.

[*Table* 3]

[12] McKenzie and Small (1997), using quarterly data from 5 US mobile telephony firms from 1993-1995, totaling 28 observations, estimated a composite cost function with subscribers as the output. They found mild decreasing returns do scale. Foreman and Beauvais (1999) using monthly data from a large panel of GTE wireless mobile market areas from 1996-1998, totaling 3.333 observations, estimated a translog cost function with subscribers and minutes of conversation as the outputs. They found mild increasing returns do scale. Parker and Roeller (1997) found increasing marginal costs, wheras Miravete and Roeller (2004) report constant marginal costs. Gagnepain and Pereira (forthcoming) found constant returns to scale.

[13] We also estimated models with random coefficients associated with price, i.e., mixed logit models. The models produced results very similar to those of the multinomial logit, and therefore are not reported. Since mobile telephony products are relatively homogeneous, the assumption of independence of irrelevant alternatives is not like to matter much. Besides, for aggregate data, the tests on the assumption of independence of irrelevant alternatives are hetroescedasticity tests. Even if hetroescedasticity is present, the estimators are consistent.

Models I and II divide mobile telephony into six products: a pre-paid card product and a contract product for each of the three mobile telephony firms, and two fixed telephony products, one for the incumbent *PTC*, and an aggregate product for the recent entrants in the fixed telephony market. The market shares were computed using the total number of minutes. The data for this model is plotted in figures (6) and (7).

[*Figure* 6]

[*Figure* 7]

Model I is a multinomial logit model. Model II is a nested logit model with two nests: **(i)** mobile telephony, and **(ii)** fixed telephony. We restricted the coefficients associated with each nest to be equal. In most cases this restriction was not binding. The values of the nest coefficients are statistically significant, and consistent with random utility maximization. Therefore, we reject the multinomial logit model, and its implied substitution patterns. The IV estimates differ substantially from the OLS estimates. The most relevant case is that of the price coefficient, which with the IV estimator assumes a value consistent with economic theory. We therefore base our calculations on the IV nested logit model.

Table 4 presents models III and IV.

[*Table* 4]

Models III and IV differ from models I and II, respectively, only in that the former split recent fixed telephony into several products, each corresponding to a firm: *Cabovisão*, *Novis*, *Oni*, and *Tele* 2. The comments made with respect to models I and II are valid also for models II and IV.

[*Table* 5]

[*Figure* 8]

Table 5 presents model V. This model includes only mobile telephony firms, each with its own two products. [14] In addition, the market in each period is the increase in minutes from the previous period. This is intended as an approximation of the minutes of the new consumers in the market, i.e., the consumers that are really choosing for the first time to use mobile telephony. The series are plotted in Figure 8.

The results are similar to those of the previous cases, namely in terms of estimate of the price coefficient. This lends some support to the diffusion process introduced earlier. A noticeable difference is that now the variable that captures network effects is no longer

[14] Not all series were stable enough to permit a disaggregated treatment in this model, and hence prepaid and subscription were joined for the exercise. The same comment is valid for the mintes of the fixed telephony entrants, which were left out.

significant. This suggests that at the current levels of the consumer base, network effects are no longer relevant for the decision to adhere to the services of a given firm. This result brings into question the usual way of capturing network effects in demand models.[15] The significance of the network size variable in a levels regression could just be capturing general dynamic elements, rather than proper network effects.

4.4 Price Elasticities of Demand

We computed the price elasticities of demand for each of the models described in the previous section. Tables 6, 7, and 8 present these values for the nested logit models II, IV, and V, respectively.

[*Table* 6]

[*Table* 7]

[*Table* 8]

Consumers have elastic demands for mobile telephony services. The demands of the fixed telephony entrants are smaller, but still elastic. The demand of *PTC* has an own-price elasticity slightly higher than 1, and in some models lower than 1.

5 Policy Analysis

Next we perform three policy exercises. First, we simulate the effect of a merger between *Tmn* and *Optimus*. Second, we simulate the effect of the entry of a firm with the characteristics of *Optimus* after the merger of *Tmn* and *Optimus*. Third, we simulate the effect of the entry of a firm with the characteristics of *Optimus* without the merger of *Tmn* and *Optimus*.

5.1 Merger of Tmn and Optimus

The merger of *Tmn* and *Optimus* would result in a market with two mobile telephony firms: **(i)** a firm controlling the products of *Tmn* and *Optimus*, and **(ii)** *Vodafone*, which would maintain its products. The merger consists of a change from matrix Γ to matrix Γ', given by:

$$\Gamma' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

[15] See, e.g., Kim and Kwon (2003), Madden and Dalzell (2004), or Okada and Hatta (1999).

We impose the assumption that marginal costs are constant. In the present context this means mostly that we can only identify point estimates of marginal costs and not the cost function itself.[17]

[*Table* 9]

Given the elasticities in Table 6, we solved system (7) numerically with respect to $\mathbf{c}$. We obtained estimates of marginal costs, $\hat{\mathbf{c}}$, which are presented in Table 9. Then, given the value of these estimates, and replacing Γ with Γ', we solved system (7) with respect to prices, to estimate the price of each product after the merger. The results are also presented in Table 9.

After the merger, the prices of mobile telephony increase on average 0.012 euros per minute, i.e., increase on average 6.3% of their current levels. The largest increases occur for *Optimus*, for which prices increase by as much as 0.033 euros per minute, i.e., increase on average 17% of their current levels.

[*Table* 10]

[*Table* 11]

The elasticities presented in Tables 7 and 8, generate smaller and larger price increases, respectively, than those of Table 6, as can be seen in Tables 10 and 11. We chose the elasticities of Table 9 only because they seem to be closer to the others found in the literature.[16]

[*Table* 12]

Table 12 reports the impact of the merger on welfare and market shares. After the merger, on average, the consumer surplus of each consumer decreases by 6.332×10^{-3} euros per minute, and profits by 6.642×10^{-3} euros per minute. Social welfare increases by 0.310×10^{-3} euros per minute. This last result is counter intuitive. However, it can be understood as consequence of some of the restrictions imposed in the demand equations. This model does not include the option to stop consuming minutes. As a consequence,

[16] We note also that the model presented in Table 4 estimates the demand of four additional products, when compared to that of Table 3, using only four extra parameters, while the number of elasticities computed increases quadratically. If it were not for the presence of the new fixed telephony products in the model presented in 3, comparison between the two models would amount to a Hausman-McFadden type test to the validity of extending the assumptions on the pattern of substitutability to the new fixed telephony products in the model of table 4. The difference between the coefficients common to both models suggests that the extension is not valid. One can also test wether a baseline model that excludes new fixed telephony can be extended, in the sense that the coefficients that characterize the utility function do not change, to the models in table 3 and to those of table 4. This extension is not rejected in the former case but is in the latter. This cold be due to the low power of the test but is nevertheless evidence against the models in table 4.

consumers cannot exit the market in response to price increases, as would be the case in a classic demand model. Instead, consumers can only switch to other lower cost, and therefore lower price firms. The effect of consumers switching to lower cost firms increases profits enough to equate or even surpass the loss in consumer surplus. If consumers where allowed to exit the market, the social welfare would likely decrease as a result of the merger.[17]

The apparent discrepancy between the magnitude of the average price decrease and the magnitude of the increase in consumer surplus is explained by the fact that the price variation captures only the welfare effect of the marginal consumer, whereas the consumer surplus also captures the welfare effect of the submarginal consumers, including those that are not directly affected by the price change. We also have that the marginal utility of income, α, is 5.9.

Plausibility of the Nash ex-ante Assumption We assumed that before the merger, firms played a Bertrand game. But firms could have played a game that led to either more or less competitive outcomes, than those implied by a Bertrand game. The marginal costs reported in Table 1, and the observed average prices, imply observed price-cost margins for *Tmn*, *Vodafone* and *Optimus* of 0.060, 0.077 and 0.063, respectively. The demand estimates and the Bertrand-Nash assumption imply estimated price-cost margins for the subscription and prepaid products of *Tmn* of 0.074 and 0.073, respectively. For *Vodafone* and *Optimus* the equivalent values are 0.067 and 0.068, and 0.051 and 0.050, respectively. Comparing these two sets of estimates, one concludes that for *Tmn* the observed margins are lower than the estimated margins. In other words, the observed behavior is more competitive than that predicted by the Nash behavior assumption. The reverse occurs for the other two firms. The variance of the parameter estimates in the cost function imply a confidence interval for the marginal costs of ± 0.023 euros, which places the estimated margins in the confidence interval of the observed ones. We interpret these results as a lack of evidence against the assumption that firms play a Bertrand game.[18] Furthermore, we discard the possibility of collusive behavior.

[17] In order to include as the outside good the option of not calling one has to have a measure of the total size of the market in minutes, including uncalled minutes. One possibility which we are exploiting but not included in the current paper is to estimate the potential number to total minutes as the saturation point of a diffusion model and use it as the market size. Alternatively one could do sensitivity analysis of the results in this paper to different assumptions of the total size of the market.

[18] Note, however, that this is not a formal test of the hypothesis since we have not taken into account the variance in the estimated margins.

5.2 Entry After the Merger

We evaluated the effect of the entry of a new firm in the mobile telephony market after[19] the merger. We assumed that this new firm would have the same characteristics as the pre-paid card product of *Optimus* in terms os prices and consumer preferences. We also computed the effect of the new firm setting prices 10% and 20% below *Optimus* prices as well, as the new Bertrand-Nash equilibria. The calculation of market shares is done using the choice probabilities for the nested logit with the parameter estimates of Table 3. The indirect utility for the new product was computed with the assumptions just mentioned above. The results in terms os market shares and consumer welfare and change in overall profits are presented in Table 12.

[*Table* 12]

A new firm that entered the market after the merger could obtain a market share of 5%, if it set prices equal to the industry post-merger average and competitors did not react to its entry. This value would be around 12% if lowered prices by 20%, and still did not have a reaction from competitors. If all firms reacted to the entry in a Bertrand-Nash fashion, the market share of the entrant would be of 10%, and it would imply lowering prices significantly, by about 17%.

These values are closer to market shares of consumers that are not yet in the market, than to market shares of consumers stolen from the other firms. Therefore, they should be taken as upper bounds of potential market shares. For the new firm, the capacity of stealing clients from the other firms is smaller than the capacity of attracting consumers that do not have yet subscribed to mobile telephony services. Since the market is almost saturated, there are few new clients to whom the market shares above apply. To the remaining clients, which are the majority, smaller market shares apply.

The consumer welfare would increase with the entry of the new firm, but not enough to compensate the effect of the merger. Profits would decrease almost symmetrically to the increase in consumer surplus, so the social welfare would increase slightly. The equilibrium with entry after the merger is less competitive than the equilibrium before the merger.

6 Concluding Remarks

In this article, we developed a model of the mobile telephony industry. We then estimated the model for a rich panel of firm level Portuguese data, and used it to perform several policy exercises.

Our consumer decision model has two components. First, the entry into the market by consumers follows a diffusion process. Second, the consumer structural decision model we

assume a discrete choice model. For the cost model we assume a quadratic cost function.

The demand model on which we base our conclusions is a nested logit model. Households are quite sensitive to price variations in mobile telephony.

We use the model to perform three policy exercises. First,we evaluated the impact of the potential merger of the first and third mobile telephony firms that would take from three to two the number of firms in the market. Our results suggest that the decrease in competition, caused by the merger, may lead to substantial price increases, as well as a decrease in consumer welfare. The merger seems to generate small efficiency gains, if any, and marginal cost reductions would have little impact on prices. Second, we evaluated the impact of the entry of a new firms after the merger. Entry would mitigate the anti-competitive effects of the merger but would not restore the pre-merger welfare levels.

Third, we evaluated the impact of the entry of a new firm without the merger.

References

Anderson, S., and A. de Palma (1992): "Multiproduct Firms: A Nested Logit Approach," *Journal of Industrial Economics*, 40(3), 261–76.

Baker, J., and T. Bresnahan (1985): "The Gains from Merger or Collusion in Product-Differentiated Industries," *Journal of Industrial Organization*, 33(4), 427–44.

Berry, S. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25(2), 242–262.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–90.

Caplin, A., and B. Nalebuff (1991): "Aggregation and Imperfect Competition: On the Existence of Equilibrium," *Econometrica*, 59(1), 25–59.

Domencich, T., and D. McFadden (1975): *Urban Travel Demand: A Behavioral Analysis*. North-Holland Publishing.

Dube, J. (2005): "Product Differentiation and Mergers in the Carbonated Soft Drink Industry," *Journal of Economics and Management Science*, 14(4), 879–904.

Foreman, R., and E. Beauvais (1999): "Scale Economies in Cellular Telephony: Size Matters," *Journal of Regulatory Economics*, 16, 297–306.

Gagnepain, G., and P. Pereira (forthcoming): "Entry, Cost Reduction, and Competition in the Portuguese Mobile Telephony Industry," *International Journal of Industrial Organization*, 25(3), 461–82.

Hausman, J. (1997): "Valuation and the Effect of Regulation on New Services in Telecommunications," *Brookings Papers in Economic Activity, Microeconomics*, pp. 1–38.

Hausman, J. (2000): "Efficiency Effects on the U.S. Economy from Wireless Taxation," *National Tax Journal*, 53(3), 733–42.

Ivaldi, M. (2005): "Study on Competition Policy in the Portuguese Insurance Sector: Econometric Measurement of Unilateral Effects in the CAIXA / BCP Merger Case," Discussion Paper 7, Autoridade da Concorrência Working Papers.

Ivaldi, M., and F. Verboven (2005): "Quantifying the Effects from Horizontal Mergers in European Competition Policy," *International Journal of Industrial Organization*, 23(9), 699–702.

J. Hausman, G. L., and J. Zona (1994): "Competitive Analysis with Differenciated Products," *Annales D'Économie et de Statistique*, 34, 159–80.

Kim, H., and N. Kwon (2003): "The Advantage of Network Size in Acquiring New Subscribers: A Conditional Logit Analysis of the Korean Mobile Telephony Market," *Information Economics and Policy*, 15(1), 17–33.

Madden, G., C.-N. G., and B. Dalzell (2004): "A Dynamic Model of Mobile Telephony Subscription Incorporating a Network Effect," *Telecommunications Policy*, 28, 133–44.

Mcfadden, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, pp. 105–42. Academic Press.

McFadden, D. (1978): "Modeling the choice of residential location," in *Spatial interaction theory and planning models*, ed. by A. Karlkvist, L. Lundkvist, F. Snickars, and J. Weibull, pp. 75–96. North-Holland, Amsterdam.

——— (1981): "Structural Discrete Probability Models Derived from Theories of Choice," in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski, and D. L. McFadden, chap. 5. Cambridge: The MIT Press.

McKenzie, D., and J. Small (1997): "Econometric Cost Structure Estimates for Cellular Telephony in the United States," *Journal of Regulatory Economics*, 12, 147–57.

Miravete, E., and L. Roeller (2004): "Competitive Nonlinear Pricing in Duopoly Equilibrium: The Early U.S. Cellular Telephone Industry," Discussion paper, University of Pennsylvania.

Nevo, A. (2000): "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *RAND Journal of Economics*, 31(3), 395–421.

——— (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69(2), 307–42.

Okada, Y., and Y. Hatta (1999): "The Interdependent Telecommunications Demand and Efficient Price Structure," *Journal of the Japanese and International Economics*, 13, 311–35.

Parker, P., and L. Roeller (1997): "Collusive Conduct in Duopolies: Multimarket Contact and Cross-Ownership in the Mobile Telephony Industry," *RAND Journal of Economics*, 28(2), 304–22.

Pereira, P., and J. Pernias (2006): "The Diffusion of Mobile Telephony in Portugal Before UMTS: A Time Series Approach," *Universidad de Valencia*. 23

Pinkse, J., and M. Slade (2004): "Mergers, Brand Competition, and the Price of a Pint," *European Economic Review*, 48(3), 617–43.

Rodini, M., W.-M., and G. Woroch (2003): "Going Mobile: Substitution Between Fixed and Mobile Access," *Telecommunications Policy*, 27, 457–76.

Small, K., and H. Rosen (1981): "Applied Welfare Economics and Discrete Choice Models," *Econometrica*, 49(1), 105–30.

7 Tables

Table 1: Cost functions I

Variable	Model 1	Model 2	Model 3	Model 4
c_{tmn}	3.773	4.528	9.753	8.774
	11.462	12.533	3.934	2.935
c_{vod}	4.167	4.904	10.204	9.189
	12.957	13.822	4.078	3.047
c_{opt}	4.060	4.910	10.156	9.226
	12.935	13.570	4.022	3.037
$\log(y)$	0.630	0.574	-0.377	-0.165
	24.258	20.994	-0.909	-0.318
$\log(y)^2$			0.041	0.030
			2.432	1.431
$\log(\omega_l)$	0.257	0.142	0.164	0.122
	3.914	1.960	2.203	1.671
$\log(\omega_k)$	0.253	0.180	0.128	0.098
	2.994	2.259	1.323	1.006
time	-0.001	0.006	-0.005	0.000
	-0.639	1.950	-1.967	0.069
time*(vod==1)		-0.003		-0.000
		-1.494		-0.184
time*(opt==1)		-0.015		-0.012
		-4.146		-2.815
N	104	104	104	104
R^2	0.972	0.976	0.974	0.977
Scale_{tmn}	0.630	0.574	0.799	0.711
Scale_{vod}	0.630	0.574	0.748	0.673
Scale_{opt}	0.630	0.574	0.690	0.630
$\text{Scale}_{tmn+opt}$	0.630	0.574	0.818	0.725
AvgC_{tmn}	0.156	0.171	0.162	0.169
AvgC_{vod}	0.299	0.308	0.297	0.310
AvgC_{opt}	0.342	0.315	0.340	0.315
$\text{AvgC}_{tmn+opt}$	0.143	0.155	0.155	0.158
MgC_{tmn}	0.099	0.098	0.129	0.120
MgC_{vod}	0.188	0.177	0.222	0.209
MgC_{opt}	0.215	0.181	0.235	0.198
$\text{MgC}_{tmn+opt}$	0.090	0.089	0.126	0.115
Q Min AvgC (10^6)	0.000	0.000	22.811	228.515
Min AvgC	0.000	0.000	0.126	0.085
Q Min AvgC^b (10^6)	0.000	0.000	2.842	3.770

Table 2: Cost functions II

Variable	Model 5	Model 6	Model 7	Model 8
c_{tmn}	3.062	3.894	8.462	1.835
	8.358	11.246	3.398	0.622
c_{vod}	3.400	4.162	8.870	2.076
	9.266	12.105	3.517	0.695
c_{opt}	3.306	4.240	8.820	2.145
	9.193	12.166	3.472	0.715
$\log(y)$	0.723	0.666	-0.184	1.024
	20.209	20.471	-0.442	2.009
$\log(y)^2$			0.036	-0.014
			2.191	-0.703
$\log(\omega_l)$	0.228	0.075	0.145	0.082
	3.660	1.147	2.029	1.240
$\log(\omega_k)$	0.207	0.063	0.114	0.094
	2.471	0.826	1.237	1.067
time	-0.003	0.006	-0.006	0.008
	-1.495	2.141	-2.519	1.814
time*(vod==1)		-0.004		-0.005
		-1.951		-1.928
time*(opt==1)		-0.020		-0.021
		-5.943		-5.127
$\log(\omega_a)$	0.127	0.184	0.108	0.195
	3.417	5.490	2.875	5.291
$\log(\omega_m)$	0.088	0.091	0.091	0.094
	3.108	3.520	3.281	3.597
N	104	104	104	104
R^2	0.976	0.982	0.977	0.982
Scale_{tmn}	0.723	0.666	0.872	0.605
Scale_{vod}	0.723	0.666	0.826	0.623
Scale_{opt}	0.723	0.666	0.774	0.644
$\text{Scale}_{tmn+opt}$	0.723	0.666	0.889	0.598
AvgC_{tmn}	0.155	0.174	0.160	0.175
AvgC_{vod}	0.296	0.300	0.298	0.299
AvgC_{opt}	0.334	0.298	0.333	0.297
$\text{AvgC}_{tmn+opt}$	0.145	0.161	0.155	0.159
MgC_{tmn}	0.112	0.116	0.139	0.106
MgC_{vod}	0.214	0.200	0.246	0.186
MgC_{opt}	0.242	0.199	0.258	0.191
$\text{MgC}_{tmn+opt}$	0.105	0.107	0.138	0.095
Min AvgC	0.000	0.000	11.151	0.000
Min AvgC	0.000	0.000	0.143	2.587
Q Min AvgC^b (10^6)	0.000	0.000	1.629	133.595

Table 3: Demand estimates I

	Model I: Multinomial Logit				Model II: Nested Logit			
	No IVs		with IVs		No IVs		with IVs	
var	coef	tstat	coef	tstat	coef	tstat	coef	tstat
price	-2.703	-1.719	-43.109	-5.783	0.806	2.826	-5.919	-3.884
network	-0.220	-3.607	-0.229	-4.085	0.098	8.316	0.091	7.803
price sms	-1.840	-0.527	11.217	2.913	-3.471	-5.555	-1.895	-2.499
time	0.093	17.844	0.050	5.579	0.047	41.627	0.042	24.473
TMN sub	-4.754	-12.728	2.323	1.752	-3.255	-46.513	-2.136	-8.186
TMN pre	-4.311	-12.294	1.942	1.651	-3.120	-48.109	-2.128	-9.171
VOD sub	-4.992	-13.214	1.663	1.325	-3.087	-42.504	-2.042	-8.333
VOD pre	-5.490	-16.248	-0.383	-0.392	-3.150	-45.819	-2.371	-12.607
OPT sub	-5.979	-10.141	-0.981	-0.925	-2.879	-25.255	-2.090	-10.319
OPT pre	-6.005	-10.436	-1.828	-1.969	-2.856	-25.513	-2.202	-12.464
OTH fix	-6.554	-26.908	-4.057	-8.125	-2.694	-38.742	-2.418	-24.644
Nest coef					0.761	71.160	0.740	66.695
R^2	0.850		0.872		0.995		0.995	
F	886.293		1073.440		32313.574		33631.493	
N	168		168		168		168	

Table 4: Demand estimates II

	Model III: Multinomial Logit				Model IV: Nested Logit			
	No IVs		with IVs		No IVs		with IVs	
var	coef	tstat	coef	tstat	coef	tstat	coef	tstat
price	6.558	3.992	-52.279	-10.842	2.296	6.824	-6.804	-2.312
network	-0.451	-6.674	-0.495	-8.623	0.190	11.834	0.202	11.722
price sms	1.462	0.318	14.440	3.516	-3.645	-3.922	-3.100	-2.771
time	0.120	24.496	0.099	22.578	0.036	24.608	0.031	14.673
TMN sub	-6.586	-15.032	2.403	3.096	-3.296	-33.453	-1.766	-3.690
TMN pre	-5.954	-14.280	1.834	2.677	-3.154	-34.307	-1.813	-4.319
VOD sub	-7.061	-16.142	1.290	1.779	-2.994	-28.977	-1.533	-3.413
VOD pre	-7.204	-18.009	-1.107	-1.976	-2.981	-30.395	-1.866	-5.501
OPT sub	-8.434	-11.647	-2.092	-2.831	-2.591	-15.681	-1.302	-3.382
OPT pre	-8.271	-11.625	-3.125	-4.598	-2.539	-15.643	-1.435	-4.327
CAB fix	-8.812	-37.221	-9.679	-45.009	-2.274	-23.083	-2.263	-20.472
NOV fix	-9.302	-38.728	-7.075	-26.945	-2.401	-23.303	-1.915	-11.896
ONI fix	-9.294	-36.854	-5.914	-17.718	-2.420	-23.294	-1.757	-8.474
TEL fix	-9.590	-33.927	-8.830	-35.984	-2.525	-23.127	-2.265	-18.100
VOD fix	-10.030	-41.816	-7.841	-30.102	-2.540	-23.105	-2.054	-12.546
Nest coef					0.809	75.918	0.905	68.921
R^2	0.896		0.925		0.996		0.995	
F	1997.315		2864.948		54181.832		46488.151	
N	246		246		246		246	

Table 5: Demand estimates III

	Model V: Multinomial Logit			
	No IVs		with IVs	
var	coef	tstat	coef	tstat
price	-21.541	-9.503	-12.489	-3.428
network	0.148	1.851	-0.039	-0.333
price sms	-12.084	-2.391	-13.719	-1.727
VOD	2.552	7.354	1.129	2.175
OPT	2.280	3.426	0.831	0.801
R^2	0.896		0.758	
F	372.395		134.729	
N	48		48	

Table 6: Elasticities I: Model II

$\frac{\partial Q_i}{\partial p_j}\frac{p_j}{Q_i}$	Fixed	TMN_{sub}	TMN_{pre}	VOD_{sub}	VOD_{pre}	OPT_{sub}	OPT_{pre}	OTH_{fix}
Fixed	-0.919	0.112	0.168	0.147	0.081	0.042	0.039	0.673
TMN_{sub}	0.129	-4.270	1.113	0.976	0.536	0.279	0.256	0.100
TMN_{pre}	0.129	0.746	-3.437	0.976	0.536	0.279	0.256	0.100
VOD_{sub}	0.129	0.746	1.113	-3.738	0.536	0.279	0.256	0.100
VOD_{pre}	0.129	0.746	1.113	0.976	-3.305	0.279	0.256	0.100
OPT_{sub}	0.129	0.746	1.113	0.976	0.536	-4.024	0.256	0.100
OPT_{pre}	0.129	0.746	1.113	0.976	0.536	0.279	-3.583	0.100
OTH_{fix}	0.870	0.112	0.168	0.147	0.081	0.042	0.039	-1.087

Table 7: Elasticities II: Model IV

$\frac{\partial Q_i}{\partial p_j}\frac{p_j}{Q_i}$	Fixed	TMN_{sub}	TMN_{pre}	VOD_{sub}	VOD_{pre}	OPT_{sub}	OPT_{pre}	CAB_{fix}	NOV_{fix}	ONI_{fix}	TEL_{fix}	VOD_{fix}
Fixed	-1.249	0.120	0.179	0.207	0.113	0.027	0.025	0.127	0.308	0.449	0.083	0.143
TMN_{sub}	0.136	-6.878	1.760	2.034	1.116	0.264	0.243	0.013	0.031	0.045	0.008	0.014
TMN_{pre}	0.136	1.180	-5.540	2.034	1.116	0.264	0.243	0.013	0.031	0.045	0.008	0.014
VOD_{sub}	0.136	1.180	1.760	-5.629	1.116	0.264	0.243	0.013	0.031	0.045	0.008	0.014
VOD_{pre}	0.136	1.180	1.760	2.034	-5.123	0.264	0.243	0.013	0.031	0.045	0.008	0.014
OPT_{sub}	0.136	1.180	1.760	2.034	1.116	-6.674	0.243	0.013	0.031	0.045	0.008	0.014
OPT_{pre}	0.136	1.180	1.760	2.034	1.116	0.264	-5.940	0.013	0.031	0.045	0.008	0.014
CAB_{fix}	1.363	0.120	0.179	0.207	0.113	0.027	0.025	-1.199	0.308	0.449	0.083	0.143
NOV_{fix}	1.363	0.120	0.179	0.207	0.113	0.027	0.025	0.127	-2.930	0.449	0.083	0.143
ONI_{fix}	1.363	0.120	0.179	0.207	0.113	0.027	0.025	0.127	0.308	-3.466	0.083	0.143
TEL_{fix}	1.363	0.120	0.179	0.207	0.113	0.027	0.025	0.127	0.308	0.449	-2.226	0.143
VOD_{fix}	1.363	0.120	0.179	0.207	0.113	0.027	0.025	0.127	0.308	0.449	0.083	-3.110

Table 8: Elasticities III: Model V

$\frac{\partial Q_i}{\partial p_j}\frac{p_j}{Q_i}$	TMN	VOD	OPT
TMN	-1.250	1.523	0.458
VOD	1.104	-2.209	0.458
OPT	1.104	1.523	-3.267

Table 9: Marginal cost estimates and post-merger prices I

Product	p_0	mc	$\frac{p_0-mc}{p_0}$	p_1	p_1^a	p_1^b	$\Delta p_1\%$	$\Delta p_1^a\%$	$\Delta p_1^b\%$	mktsh
Fixed	0.078	-0.007	1.089	0.079	0.079	0.079	0.387	0.285	0.138	0.277
TMN_{sub}	0.220	0.148	0.329	0.231	0.226	0.221	5.001	2.660	0.417	0.086
TMN_{pre}	0.200	0.127	0.362	0.211	0.207	0.203	5.555	3.447	1.505	0.142
VOD_{sub}	0.207	0.140	0.325	0.212	0.211	0.209	2.625	1.786	1.084	0.120
VOD_{pre}	0.169	0.101	0.399	0.174	0.172	0.171	3.144	2.189	1.324	0.081
OPT_{sub}	0.189	0.138	0.268	0.222	0.217	0.212	17.332	14.854	12.486	0.038
OPT_{pre}	0.168	0.118	0.300	0.202	0.197	0.196	19.674	17.106	16.166	0.039
OTH_{fix}	0.077	0.006	0.920	0.077	0.077	0.077	0.293	0.235	0.096	0.218
Avg mob	0.197	0.130	0.342	0.209	0.205	0.202	6.337	4.612	3.131	0.102

p_0 - prices today; p_1 - prices after merger; p_1^a - prices after merger considering a 5% reduction in costs; p_1^b - prices after merger considering a 5% reduction in costs; mc - marginal costs

Table 10: Marginal cost estimates and post-merger prices II

Product	p_0	mc	$\frac{p_0-mc}{p_0}$	p_1	p_1^a	p_1^b	$\Delta p_1\%$	$\Delta p_1^a\%$	$\Delta p_1^b\%$	mktsh
Fixed	0.070	0.014	0.801	0.070	0.070	0.070	0.031	0.009	-0.081	0.285
TMN_{sub}	0.217	0.173	0.203	0.221	0.215	0.208	1.746	-1.100	-3.871	0.081
TMN_{pre}	0.196	0.153	0.224	0.200	0.195	0.190	1.927	-0.693	-3.233	0.134
VOD_{sub}	0.206	0.158	0.235	0.209	0.206	0.204	1.183	0.074	-1.000	0.147
VOD_{pre}	0.168	0.120	0.288	0.170	0.168	0.166	1.451	0.066	-1.230	0.099
OPT_{sub}	0.187	0.158	0.156	0.205	0.200	0.195	9.657	7.019	4.202	0.021
OPT_{pre}	0.166	0.137	0.175	0.185	0.180	0.176	10.869	8.390	5.899	0.022
CAB_{fix}	0.036	0.006	0.839	0.036	0.036	0.036	-0.506	-0.109	-0.457	0.052
NOV_{fix}	0.087	0.057	0.341	0.087	0.087	0.087	0.013	-0.034	0.013	0.052
ONI_{fix}	0.105	0.075	0.289	0.105	0.105	0.105	-0.007	0.025	-0.004	0.063
TEL_{fix}	0.062	0.034	0.450	0.062	0.063	0.062	-0.399	0.588	-0.059	0.020
VOD_{fix}	0.088	0.059	0.322	0.088	0.087	0.088	0.245	-0.285	-0.054	0.024
Avg mob	0.195	0.150	0.231	0.200	0.196	0.192	2.297	0.330	-1.584	0.113

p_0 - prices today; p_1 - prices after merger; p_1^a - prices after merger considering a 5% reduction in costs; p_1^b - prices after merger considering a 5% reduction in costs; mc - marginal costs

Table 11: Marginal cost estimates and post-merger prices III

Product	p_0	mc	$\frac{p_0-mc}{p_0}$	p_1	p_1^a	p_1^b	$\Delta p_1\%$	$\Delta p_1^a\%$	$\Delta p_1^b\%$	mktsh
TMN	0.189	0.038	0.800	0.211	0.210	0.209	11.815	11.394	11.018	0.469
VOD	0.299	0.164	0.453	0.313	0.312	0.311	4.595	4.322	4.045	0.408
OPT	0.298	0.207	0.306	0.380	0.371	0.362	27.448	24.337	21.260	0.123
Avg mob	0.247	0.110	0.597	0.273	0.271	0.269	10.789	10.098	9.430	0.402

p_0 - prices today; p_1 - prices after merger; p_1^a - prices after merger considering a 5% reduction in costs; p_1^b - prices after merger considering a 5% reduction in costs; mc - marginal costs

Table 12: Merger and new product effects on shares and CS

Product	Today	After Merger	New product I	II	III	IV
All						
PT_{fix}	0.347	0.360	0.357	0.356	0.354	0.299
OTH_{fix}	0.123	0.128	0.127	0.126	0.126	0.167
TMN_{sub}	0.105	0.104	0.099	0.097	0.093	0.094
TMN_{pre}	0.150	0.148	0.141	0.138	0.133	0.134
VOD_{sub}	0.102	0.115	0.110	0.107	0.103	0.114
VOD_{pre}	0.082	0.092	0.088	0.086	0.082	0.091
OPT_{sub}	0.045	0.027	0.026	0.025	0.024	0.024
OPT_{pre}	0.046	0.027	0.026	0.025	0.024	0.025
NEW	0.000	0.000	0.026	0.040	0.061	0.052
Mobile						
TMN_{sub}	0.198	0.202	0.192	0.187	0.178	0.176
TMN_{pre}	0.283	0.289	0.274	0.266	0.255	0.251
VOD_{sub}	0.193	0.224	0.213	0.207	0.198	0.213
VOD_{pre}	0.154	0.180	0.170	0.166	0.158	0.170
OPT_{sub}	0.084	0.052	0.050	0.048	0.046	0.045
OPT_{pre}	0.086	0.053	0.051	0.049	0.047	0.047
NEW	0.000	0.000	0.051	0.078	0.118	0.098
$\Delta\text{CS}^\dagger$	0.000	-6.332	-5.162	-4.503	-3.493	-4.610
$\Delta\pi^\dagger$	0.000	6.642	6.752	6.000	4.428	4.466

I - $p_{NEW} = p_{OPT}$; II - $p_{NEW} = 0.9\,p_{OPT}$; III - $p_{NEW} = 0.8\,p_{OPT}$; IV - New Nash eq.; † - 10^{-3} euros per minute; Variations calculated with respect to column *Today*; Total minutes per quarter: 5.5863×10^9

8 Figures

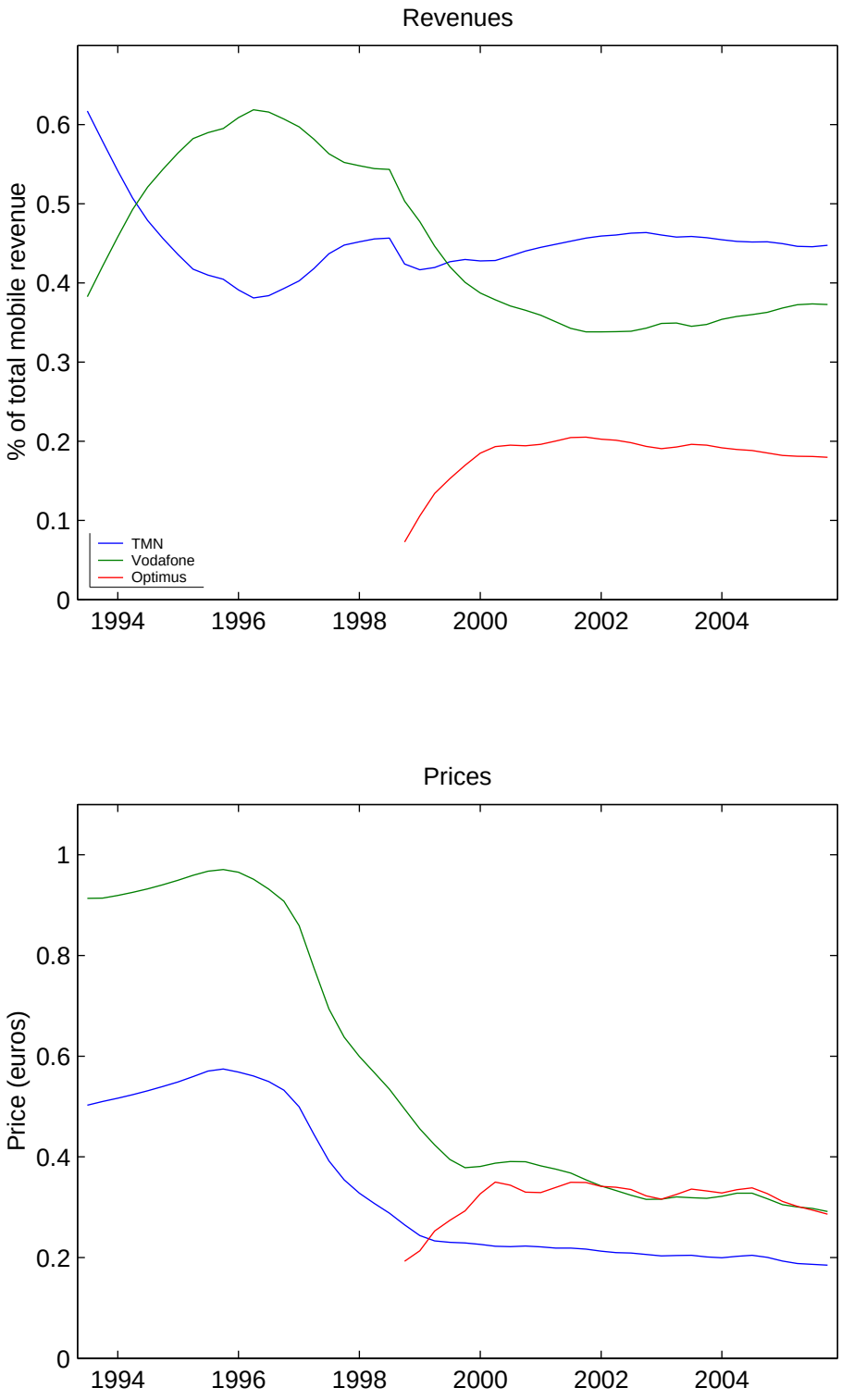


Figure 1: Mobile shares and prices

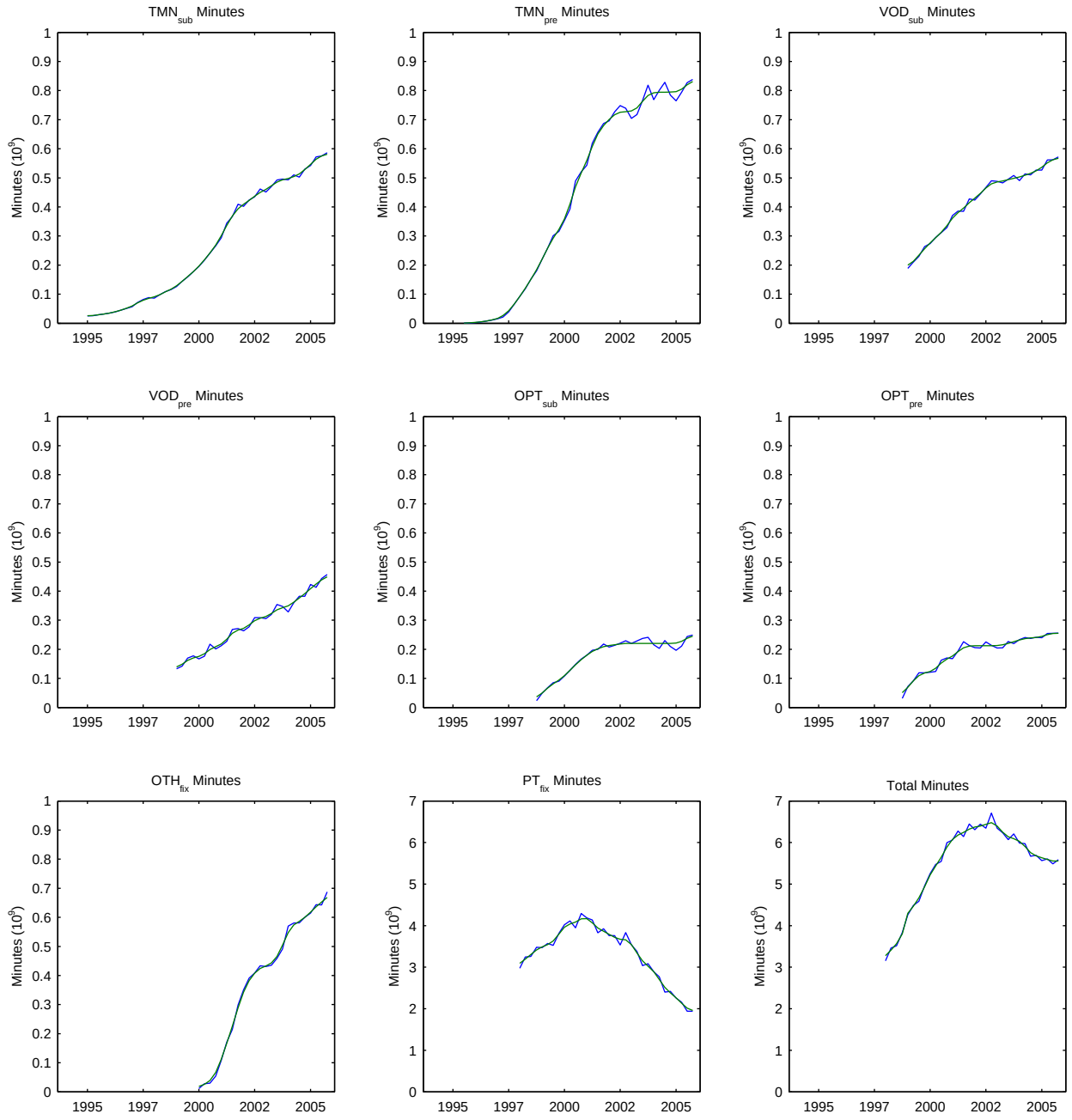


Figure 2: Observed and smoothed minutes

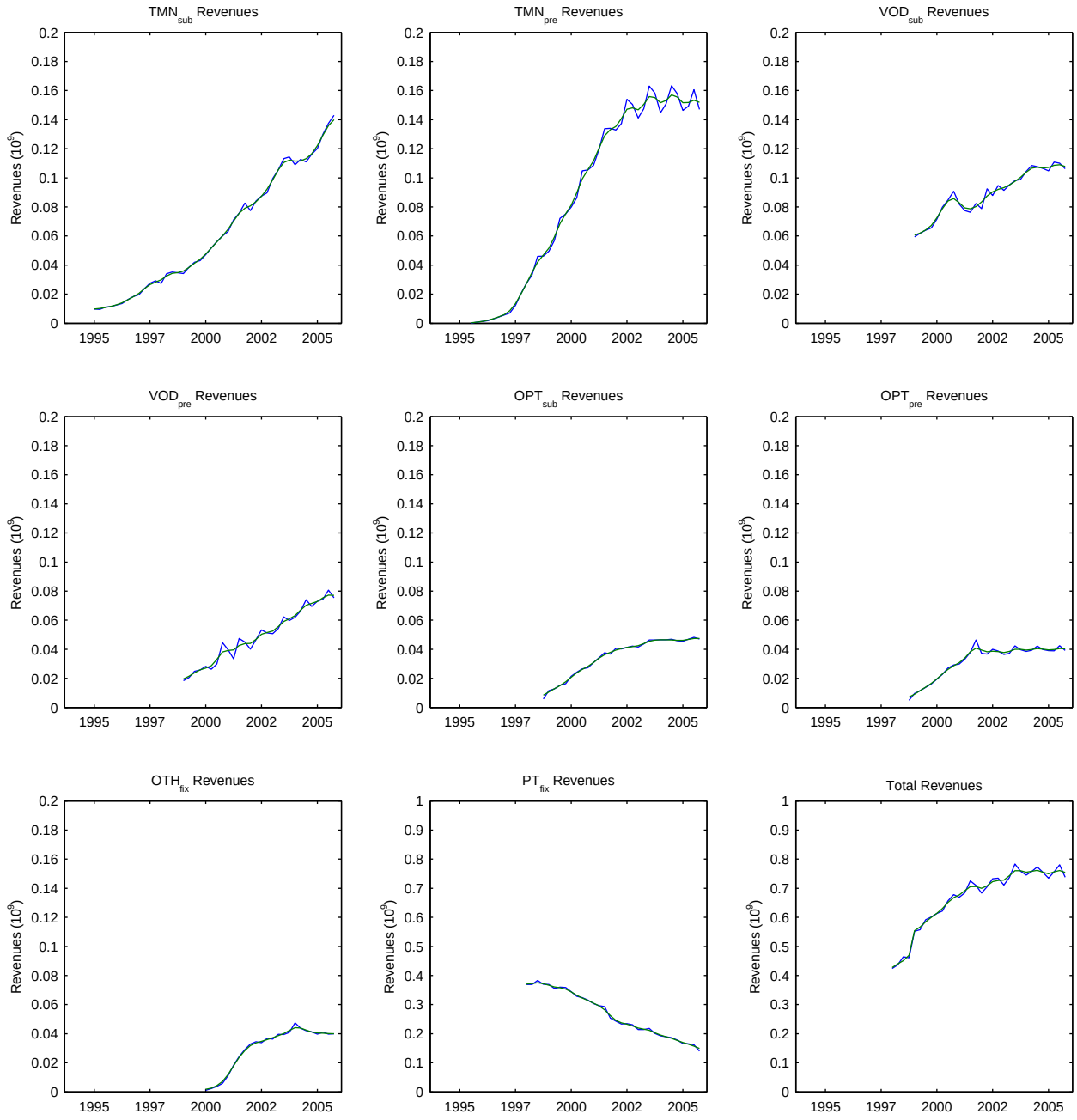


Figure 3: Observed and smoothed revenues

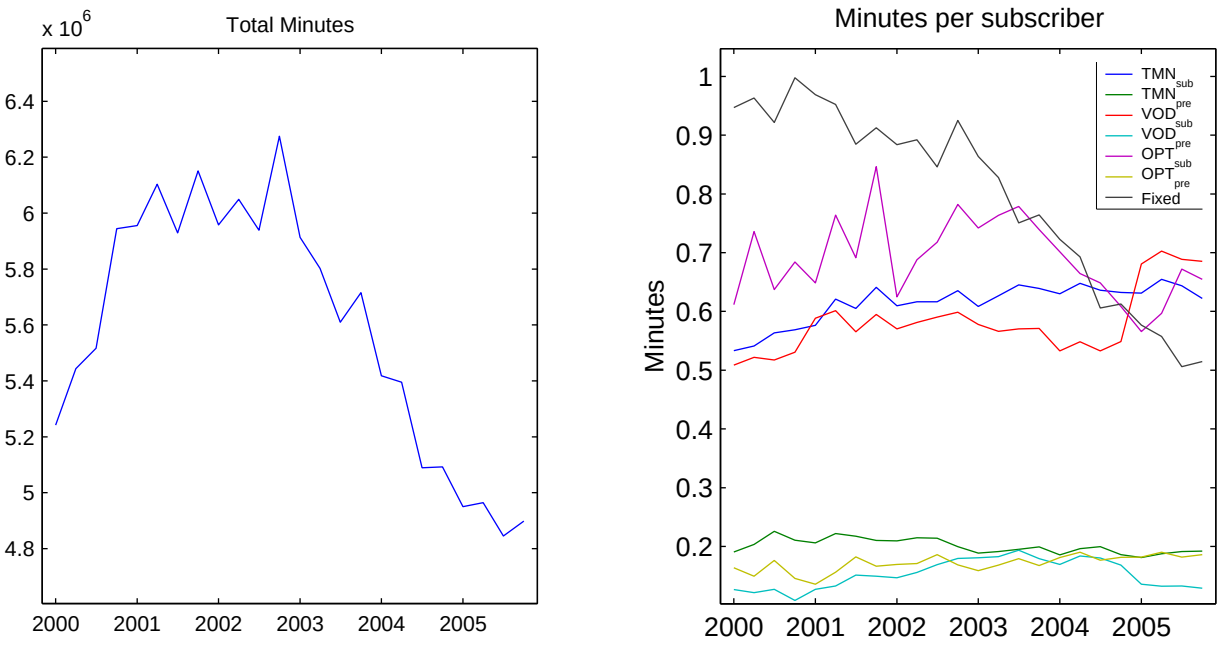


Figure 4: Minutes per subscriber

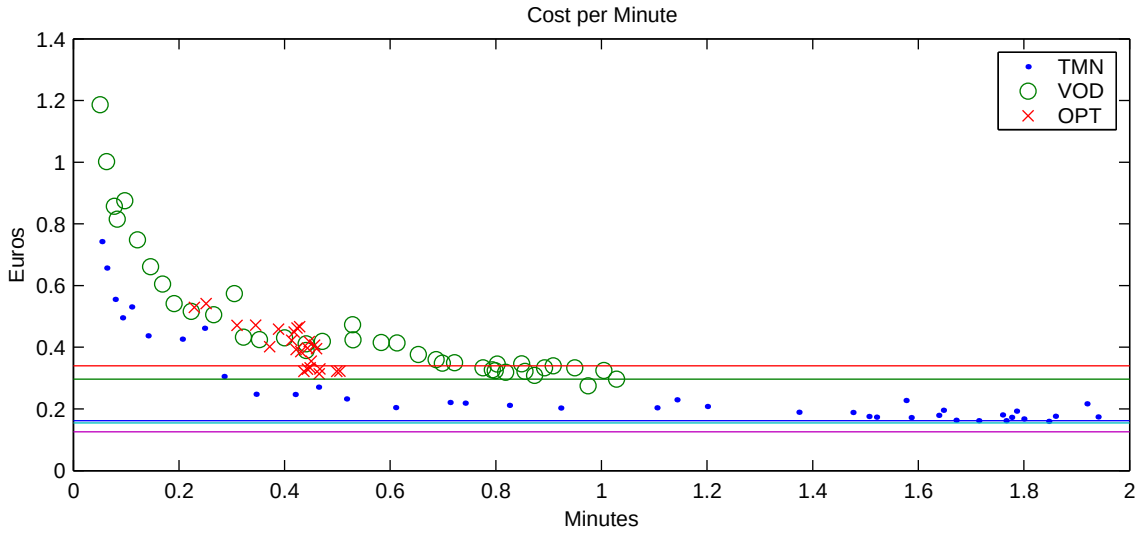


Figure 5: Average costs

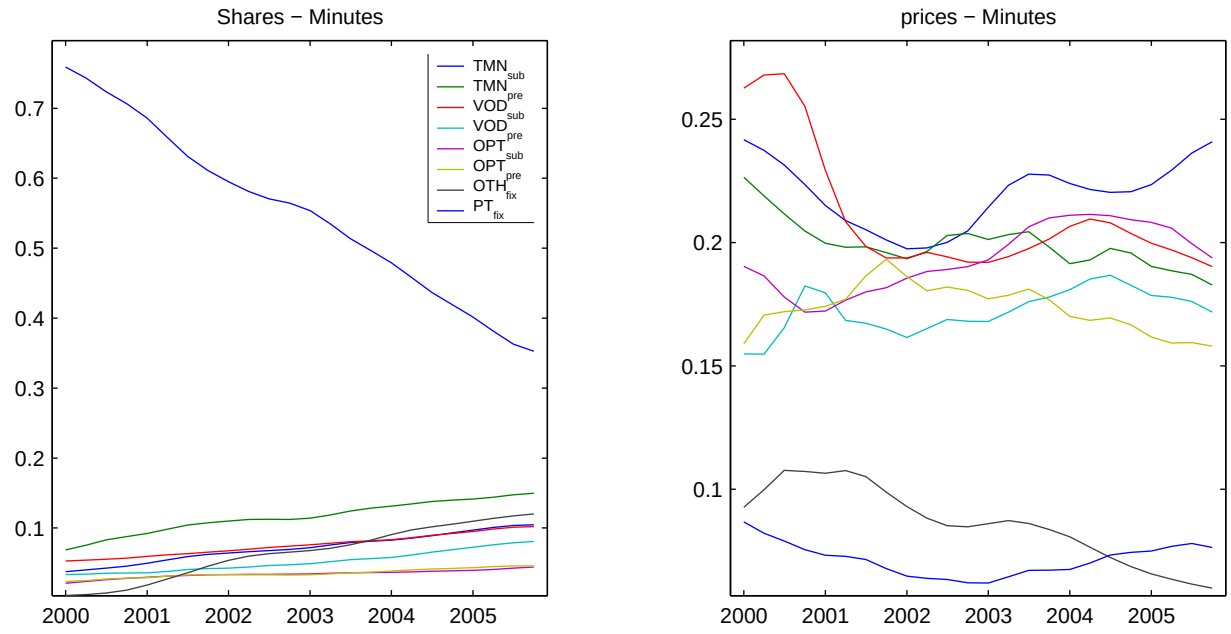


Figure 6: Shares and prices

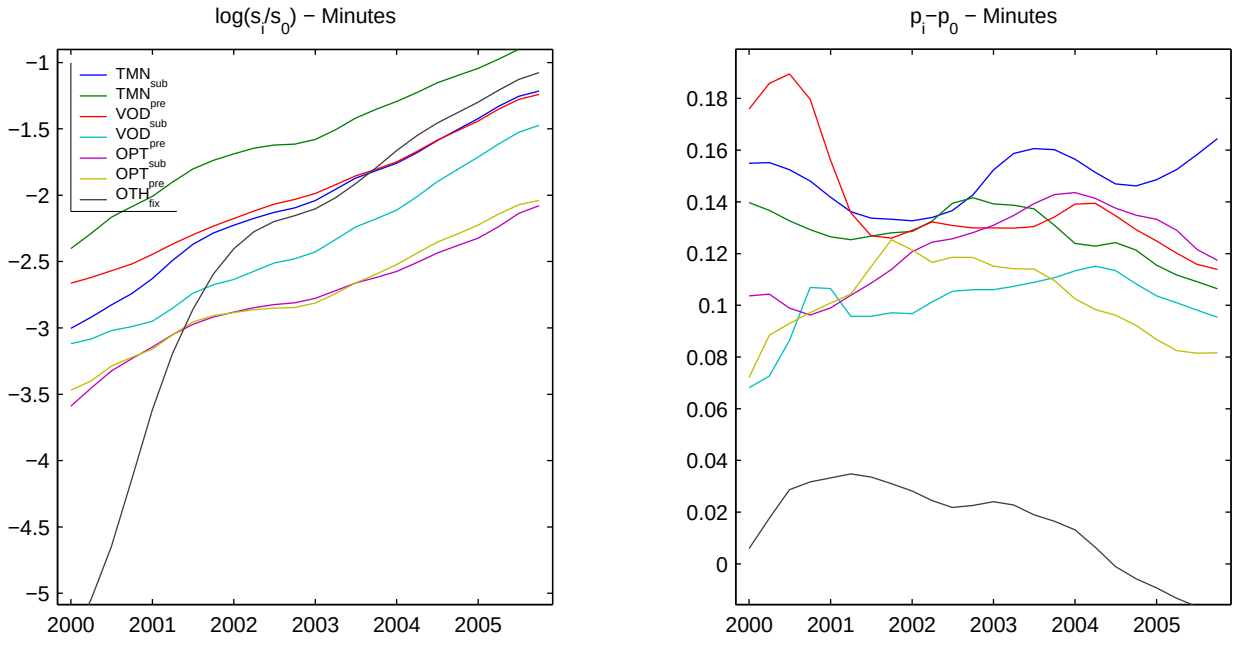


Figure 7: Transformed data

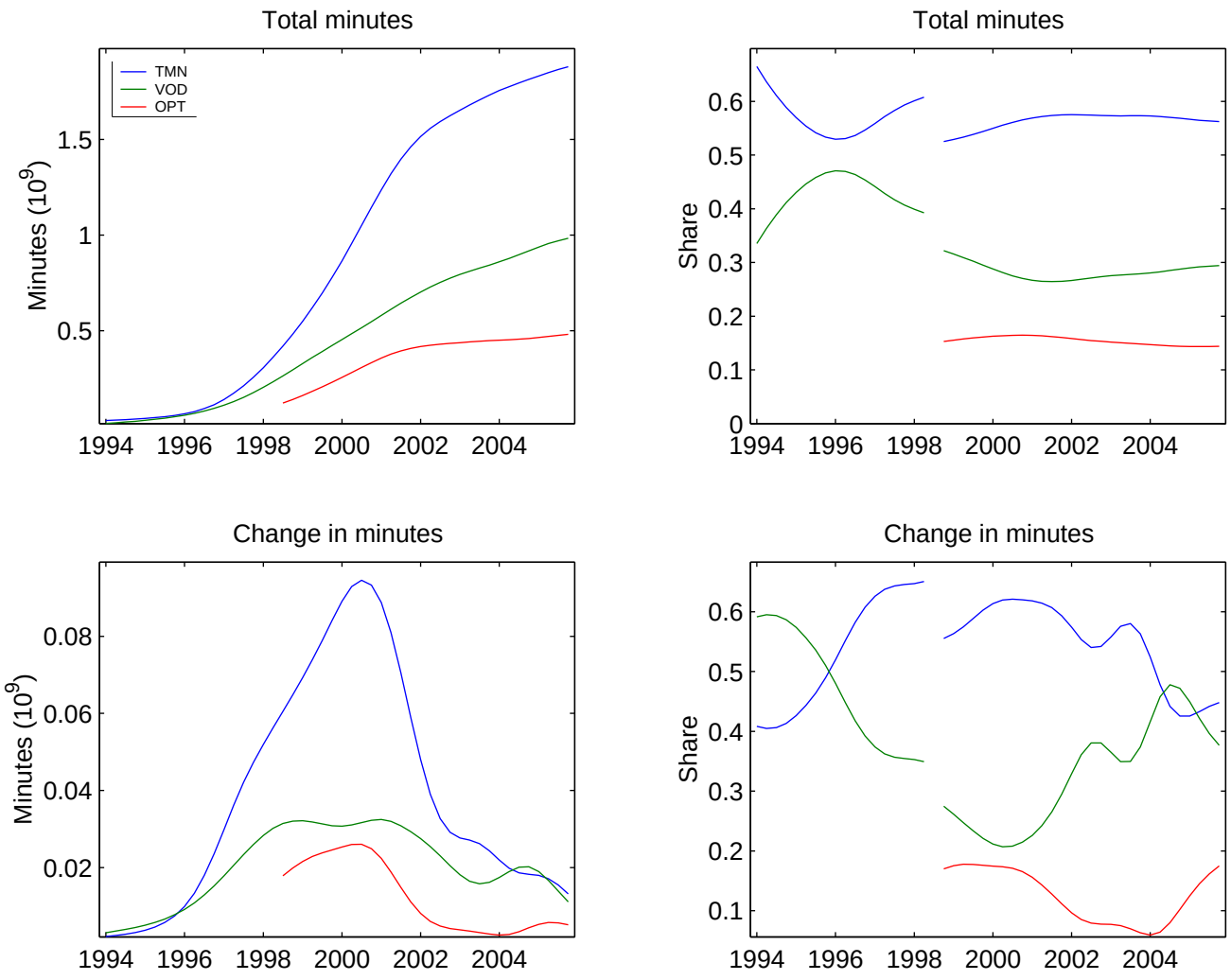


Figure 8: Change in minutes

Copyright Law, Movie Production, and Video Pricing: the European Rental Directive

I.P.L. Png and Qiu-hong Wang*

November 2007

Abstract

In November 1992, the European Economic Community issued Directive 92/100/EEC (the "Rental Directive") to harmonize copyright laws with regard to rights of rental and lending, and the neighboring rights of performers, music and film producers, and broadcasters. Member countries of the European Economic Area were required to comply with the Directive with effect from July 1994.

We studied the impact of the Rental Directive on the production of movies in 17 European countries during the period 1993-2005.

We found that the Rental Directive was associated with an increase in movie production ranging between 2.46% (±1.55%) and 4.43% (±2.76%). Importantly, the increase was higher in countries where piracy was lower. These findings were robust to various specifications, including the measure of compliance with the Rental Directive and the measure of piracy, changes in government funding, other significant changes in copyright law, and exclusion of a possible outlier country.

The Rental Directive enabled movie producers to directly discriminate between video tapes for sale vis-à-vis rental. It was associated with retail rental revenues being 0.85% higher and retail sale revenues being 1.81% higher.

* National University of Singapore. Corresponding author: Ivan Png, Dept of Information Systems, National University of Singapore, 3 Science Drive 2, Singapore 117543, Tel: +65 6516-6807, http://www.comp.nus.edu.sg/~ipng/. We thank participants at … seminars at the Hong Kong University of Science and Technology, Oxford University, … for helpful advice. Ivan Png thanks Nuffield College for kind hospitality in Michaelmas Term, 2007, during which part of this paper was written.

1. Introduction

Generally, copyright law must strike a delicate balance between two considerations:

- Broader and longer protection increases the return to creators of new work, and in the long term, encourages more creative work;
- Narrower and shorter protection increases the use of existing creative work, and hence, raises the benefit to end-users and also facilitates new creations that build upon earlier work. [1]

There is no disagreement about the directions of these two considerations (Plant 1934; Nordhaus 1969; Gallini and Scotchmer 2002). However, debate on the trade-off has been controversial. Some scholars argue that the scope and term of copyright are excessive (Lessig 2001; Boldrin and Levine 2002; Quah 2002). Others argue in favor of more protection (Landes and Posner 1989; Miller 1995).

The controversies continue to rage in part because there has been very little systematic empirical evidence on either the long-term impact of copyright on the creation of new work or the short-term impact on the use and re-use of existing work. "Perhaps the most pressing area in which the economics of copyright is lacking is in serious empirical studies" (Watt 2004).

In this paper, we study the impact of European Economic Community (EEC) Directive 92/100/EEC, the so-called "Rental Directive", on production of movies and pricing of videos. This Directive was part of the EEC's effort to establish a single European market. The Directive aimed to harmonize copyright laws in the member countries with regard to rights of rental and lending, and the neighboring rights of performers, music and film producers, and broadcasters with effect from July 1994.[2]

Prior to the Directive, copyright laws in the member countries varied in their treatment of rental and lending rights, and the various neighboring rights. We investigated the impact of the Rental Directive on the production of movies in 17 European countries during the period 1993-2005. We found that, on average, revision of national laws to

[1] An alternative is to replace intellectual property rights with a system of rewards for inventors and creators (Shavell and van Ypersele 2001).

[2] In the civil law tradition, authors have copyright, while performers, music producers and broadcasters have "neighboring rights". The common law tradition does not make such a distinction.

comply with the Directive was associated with an increase in movie production ranging between 2.46% (±1.55%) and 4.43% (±2.76%). Importantly, the increase in production was higher in countries with lower rates of piracy.

These findings were robust to exclusion of a possible outlier country and various specifications, including an alternative measure of compliance with the Rental Directive and an alternative measure of piracy, other contemporaneous changes in copyright law, and changes in government funding.

The Rental Directive allowed movie studios to directly discriminate in the sale of video tapes to retailers between those for sale vis-à-vis rental to the end-consumer. Indeed, we found that compliance with the Directive was associated with 3.86%(±0.95%) higher rental rates and 1.23%(±1.21%) lower sell-through prices. Overall, retail rental revenues and retail sale revenues were 0.85% and 1.81% higher respectively.

Our findings confirm that changes in the depth and scope of copyright law did have economically significant effects on the production and pricing of at least one category of creative work – movies.

2. Previous Research

Surprisingly, despite persistent controversy, there has been little empirical investigation of the impact of copyright law on the production of creative work (Watt 2004; Png 2006). The little extant work mostly provides only indirect evidence. For most of the 19th century, U.S. copyright law did not protect British authors. Then, in 1891, Congress passed the International Copyright Act, which extended copyright protection to foreign authors, and through reciprocal recognition, extended international copyright protection to U.S. authors. However, passage of the Act did not substantially affect the number of full-time authors in the United States (Khan 2004).

In 1998, the United States followed and passed the Sonny Bono Copyright Term Extension Act which extended the term of copyright from author's life plus 50 years to author's life plus 70 years. Hui and Png (2002) found that the Act had a positive but insignificant effect on U.S. production of movies. However, consultants to the motion picture industry criticized this study on two grounds: "relies upon such a small sample (11 years), with only two after the extension" and "ignores the significant lead time that movies require before production, and hence is likely to understate the incentives in the initial

years after extension" (Allen Consulting Group (2003), page 27).

Landes and Posner (2003) studied the impact of the 1962 extension of copyright term and the 1998 CTEA on all U.S. copyright registrations between 1910 and 2000. Three categories accounted for 70 percent of all registrations with the Copyright Office – books, music, and graphic arts. They found that both changes had positive but insignificant effects: "It is not surprising that the term-extension variables (in 1962 and 1998) are insignificant; the expected commercial life of a copyrighted work is so much shorter than the copyright term that it makes a lengthening of the term irrelevant to most potential registrants" (Landes and Posner (2003), page 247).

Reynolds (2003) used ordinary least squares to test the impact of national copyright law on the number of movies produced, as measured by submissions to the Cannes Film Festival between 1965 and 2002. While the number of movies was positively and significantly related to his index of copyright law in some specifications, it was not significantly related to the duration of copyright protection.

Reynolds' results should be interpreted with caution. The number of submissions to Cannes might not be representative of total movie production. Indeed, the one explanatory variable that was robust across all specifications was the total number of films accepted by the Festival. Moreover, the mean of the dependent variable was 0.34 and standard deviation was 1.60, suggesting that the distribution comprised many zeroes with a few positive integers. With a count dependent variable, the usual ordinary least squares test statistics are not valid, and it would be more appropriate to apply a Poisson regression (Wooldridge (2006), pp. 604-609).

Baker and Cunningham (2005) have provided the main empirical evidence of the incentive effect of copyright law. They found that court decisions broadening copyright protection were associated with increases in copyright applications between 1994-2005 in Canada and between 1986-2004 in the United States. In addition, copyright applications were higher when the economic growth was slower, which is consistent with creative activity being complementary with leisure.

The findings of Landes and Posner (2003) and Baker and Cunningham (2005) are subject to a significant limitation. They do not show the impact of copyright law on the quantity of books, movies, and music as such. Copyright registration is not compulsory. To the extent that creators of unregistered work expect lower returns (which do not justify

the cost of registration), they might be at the margin that would respond to changes in the depth and scope of copyright law.

The U.S. Copyright Term Extension Act followed the European Directive 93/98/CEE to harmonize the term of copyright to author's life plus 70 years with effect from July 1, 1995. Using a panel of [..] countries, Png and Wang (2007) found that copyright term extension was associated with …

Besides the contributions of Baker and Cunningham (2005) and Png and Wang (2007), the impact of copyright protection generally on the incentive to create new work continues to be an open question.

Empirical research into impact of copyright law on the pricing of copyrighted items is also quite sparse. Liebowitz (1985) observed that, following the widespread adoption of photocopying machines, journal publishers raised subscription rates to libraries relative to rates for individuals. Further, the differential was highest for the most frequently copied journals. By charging discriminatory rates, the publishers could "indirectly appropriate" some of the libraries' benefit from copying.

In a very sophisticated study, Mortimer (2007) estimated the retail demand for rental vis-a-vis sell through videotapes and DVDs. She calculated that direct discrimination would benefit studios and consumers at the expense of retailers in the distribution of DVDs, but not necessarily for videotapes.

3. Context

On November 19, 1992, the European Economic Community (EEC) Council of Ministers issued Directive 92/100/EEC to harmonize copyright laws with regard to rental and lending, and neighboring rights for performers, music and film producers, and broadcasters with effect from July 1, 1994.[3] This so-called "Rental Directive" was just one of a series of Directives issued to bring about a single European market. By the European Union Treaty ("Maastricht Treaty"), the EEC became the European Union (EU) with effect from November 1, 1993.

[3] Council Directive 92/100/EEC of 19 November 1992 on rental right and lending right and on certain rights related to copyright in the field of intellectual property, *O.J.* No. L 346 of 27 November 1992, 616-66. "Neighboring rights" are the European name for the rights of creators, such as performers, music and movie producers, and broadcasters, who are not authors.

Prior to the issuance of the Rental Directive, copyright law in European member countries differed in whether creators of works could control rental and lending. Differences between the copyright laws of Denmark and U.K. came to a head in the *Warner-Metronom*e case.[4] A Danish national bought video-tapes in the U.K. and offered them for rental in Denmark. The producer of the video-tapes sued to control rental of the tapes. Under Danish law, the producer could control rental, but not under U.K. law.[5]

Also, prior to the Directive, copyright law in European member countries differed in the scope of "neighboring rights" of creators other than authors. In particular, the Rome Convention provided neighboring rights to performers, music producers, and broadcasters, but not movie producers (Geller (1999) Section 4[2][c][ii]).

The key changes required by Directive 92/100/EEC were:

- Article 1: exclusive rental and lending rights;
- Article 2: director of audiovisual work to be an author, presumption of transfer of rights from performers to audio-visual producers, optional presumption of transfer of rights from authors to audio-visual producers;
- Article 4: author and performer to have unwaivable right to equitable remuneration from rental;
- Article 5: exception from exclusive lending right;
- Articles 6-9: (neighboring) rights of fixation, reproduction, broadcasting and communication to the public, and distribution for performers, music and movie producers, and broadcasters.

Based on the survey by Reinbothe and von Lewinksi (1993), we compiled in Table 1 the compliance of existing national copyright law among members of the European Union and European Economic Area (EEA) with Articles 1, 2, 4, 5, and 6-9 of the Rental Directive. By the European Economic Area Agreement, member-countries Austria, Finland, Sweden, Norway, Iceland, and Liechtenstein committed to harmonize their laws with those of the EU.[6] The EEA Agreement took effect in 1994. Although an EEA

---

[4] Judgment of 17 May 1988, Case 158/86, Warner Brothers Inc. and Metronome Video Aps. v. Erik Viuff Christiansen, [1988] E.C.R. 2605.

[5] U.S. law does not allow a video-tape producer to control rental. Mortimer (2007) estimated the impact of such control on U.S. consumer welfare, and movie producer and retailer profits.

[6] "European Economic Area – Overview", http://ec.europa.eu/external_relations/eea/index.htm [Accessed, August 4, 2007]. Subsequently, in January 1995, Austria, Finland, and Sweden joined the European Union. Switzerland did not join the EEA Agreement.

member, Switzerland decided not to accede to the EEA Agreement.

By our own further legal research, we tabulated the compliance of national copyright law in three countries – the Czech Republic, Hungary, and Poland – that subsequently joined the European Union.[7]

-- Table 1: Compliance with Rental Directive --

Where relevant, Table 1 focuses on the changes from the viewpoint of movie producers. An entry "1" indicates that the national copyright law complied with the corresponding Article of the Rental Directive. Where the national law did not comply with the Directive (as indicated by "0"), the national law had to be revised. As Table 1 shows, the survey of Reinbothe and von Lewinksi (1993) and our own legal research was incomplete.

For the effective dates of the revisions of the national law to comply with the Rental Directive, we relied on a study by the European Commission (undated) and our own legal research. According to the European Commission (undated), as of 1999, Ireland had not complied with the Rental Directive. Ireland passed the relevant legislation in 2000, but the effective date was not clear. To comply with the Directive, the law should have been effective in 1994. Accordingly, we excluded Ireland from our study.

Following several revisions, the original Rental Directive (92/100/EEC) was superseded by Directive 2006/115/EC, issued on December 12, 2006.[8]

4. Data and Specification

Copyrightable works include books, illustrations, photographs, sound recordings, audio-visual works, and software. Among these, so far as we are aware, audio-visual works is the only category about which there is comprehensive international information over a reasonable period of time. This is available from the Euromonitor's Global Market Information Database (GMID). The GMID provides information about the number of movies produced by country and year. [9]

[7] Our own legal research was based on Geller (1999) and the online collection of copyright laws provided by the World Intellectual Property Organization (http://www.wipo.int/clea/en/index.jsp).
[8] *O.J.* No. L 376 of 27 December 2006, 28-35.

[9] Two other sources are the Internet Movie Database ("IMDb"), published by Amazon.com, and the Film Index International, published by the British Film Institute. Information in IMBb is

In this study, we focused on the number of movies produced by country-year. We applied a "difference in differences" strategy, which specified movie production in country *i* and year *t* as

$$\log(\text{MOVIES}_{it}) = f(\text{DIRECTIVE}_{i}, X_{it}), \qquad (1)$$

where DIRECTIVE_{it} was an measure of compliance with the Rental Directive in country *i* and year *t*, and X_{it} was a vector of other variables that might possibly affect movie production.

The "difference in differences" specification accounted for any general changes in market or technological conditions that might possibly have affected the incentive to produce movies when the Directive took effect (Bertrand et al. 2004).

Referring to Table 1, the survey of Reinbothe and von Lewinksi (1993) and our own legal research was incomplete for compliance with Articles 5, 6, and 8 of the Rental Directive. Intuitively, the public lending right (Article 5) and right of first fixation for performers (Article 6) seemed relatively unimportant to movie producers. Further, Article 8 applied to performers, music producers, and broadcasters only (Reinbothe and von Lewinski (1993), page 92). Hence, we disregarded these Articles.

With regard to the other indicators of compliance – with Articles 1 (rental), 1 (lending), 3 (presumption of transfer), 4 (unwaivable right of remuneration), 7 (reproduction), and 9 (distribution), our information on compliance was almost complete. However, as reported in Table 2, the indicators were highly collinear. Rather than omit particular indicators, we applied principal components analysis to generate one composite measure of compliance from the six indicators.[10]

-- Table 2: Correlations in compliance --

Referring to Table 3, for each country and year, we obtained information from various sources about other national characteristics that might possibly affect the demand for movies or cost of movie production, and hence movie production – population, GDP

---

organized by title rather than country of production, and is submitted by industry members and website visitors. Moul and Shugan (2005) and Waterman and Lu (2005) used data from the IMDb. The Film Index International allows query of only one title at a time, and prohibits automated extraction of data.

[10] We checked the robustness of this approach by using the alternative of building a composite indicator by simply adding the six compliance indicators. The results were similar.

per capita, computer ownership and internet access, real interest rates, and piracy. Among these other variables, to minimize multi-collinearity, all national aggregates other than population were specified on a per capita basis.[11]

-- Table 3: Descriptive statistics --

An immediate concern was a secular trend in the movie industry towards more international co-production, as illustrated by Figure 1, which shows the [average number of co-producing countries per movie] over the period 1993-2005. To account for this trend, we disregarded movies involving co-production and focused on national productions.

-- Figure 1: [Average number of co-producing countries per movie] –

To provide an overview of the impact of the Rental Directive, Figure 2 illustrates, for each country, movie production and the degree of compliance with the Directive over the period 1993-2005. The graphs suggest a slight increase in movie production over the years. This was correlated with the increase in compliance with the Rental Directive. However, the increase in movie production might be explained by general economic growth or a fall in real interest rates.

-- Figure 2: Movie production and compliance with Rental Directive –

5. Results

We first estimated a very simple specification, regressing movie production on just country indicators using ordinary least squares. We report the results in Table 4, column (a).

In the next specification, we included the measure of compliance with the Rental Directive, as well as various demographic and financial characteristics – GDP per capita, population, computer ownership and internet access, the real interest rate, and year indicators. Table 4, column (b), reports the results. The coefficients of GDP per capita, population, and real interest rate had the expected signs but were imprecisely estimated.

The coefficient of the compliance indicator was positive and significant. An instructive measure of the impact of the Rental Directive is the effect of increasing from zero to 100% compliance with the Directive. Based on the mean number of movies produced, the increase in movie production associated with compliance was +5.10%

[11] Unless otherwise stated, all variables other than indicators were specified in natural logarithms.

(±2.07%).

In the following specification, we included a measure of the enforcement of copyright law. The movie industry has vigorously asserted that: "Film theft has an enormous impact on filmmakers everywhere … jeopardizing the creative process and robbing local economies of the benefits derived from having a healthy film industry" (Motion Picture Association of America (MPAA) 2006). Accordingly, movie production should be lower in countries where piracy is higher.

Besides directly affecting movie production, piracy should also have an indirect effect through changes in copyright law to comply with the Rental Directive. Specifically, in countries where piracy is higher, changes in copyright law should have a smaller effect on movie production.

Unfortunately, we were unable to procure data on movie piracy from the MPAA or elsewhere. However, we did manage to obtain music CD piracy rates from the International Federation of the Phonographic Industry (IFPI) for some years. Table 4, column (c), reports the results with two additional variables – the music CD piracy rate and the interaction of the measure of compliance with the Rental Directive and piracy rate (both in absolute, not logarithmic form). As expected, the coefficient of piracy was negative but not statistically significant.

Consistent with prediction, the coefficient of the interaction between the measure of compliance with the Directive and piracy rate was negative and precisely estimated. This result is quite compelling: apart from the incentive effect of copyright law, there seems to be no other reason why changes in copyright law should have smaller effects in countries with higher piracy.

Based on the coefficients of piracy and the interaction variable from specification (c) and the mean movie production and piracy rate, we calculated the impact of increasing from zero to 100% compliance with the Rental Directive on movie production to be +3.97% (±2.09%).

A potentially serious issue in difference-in-difference studies is serial correlation. This could result in standard errors being substantially under-estimated (Bertrand et al. 2004). Indeed, using a Wald test (Wooldridge 2002; Drukker 2003), the null hypothesis of no first-order serial correlation was rejected ($F = 22.82$, $\Pr(F > 22.82) = 0.0002$). In addition, we found strong evidence of heteroscedasticity in the residuals ($\chi^2 = 109.9$, $\Pr(\chi^2$

> 109.9) = 0.0000). We ruled out cross-sectional dependence. Using a […] test (Friedman …), the null hypothesis of [..] could not be rejected (χ^2 = 1.65, Pr(χ^2 > 1.65) = 1.000).[12] In subsequent specifications, we applied various methods to account for serial correlation and heteroscedasticity.

As a baseline for the estimates with adjustment for serial correlation and heteroscedasticity, we re-estimated the specification (c) using fixed-effects. Table 4, column (d), reports the results. Next, as recommended by Bertrand et al. (2004), we used fixed-effects with a robust cluster variance matrix. Table 4, column (e), reports the results. As expected, the estimated standard errors of the coefficients of the Directive and its interaction with piracy were larger than in the baseline. Accordingly, the estimated standard error of the impact of compliance with the Rental Directive on movie production was also larger. The estimated impact was +3.97% (±2.85%).

An alternative way to account for serial correlation and heteroscedasticity is to apply feasible generalized least squares (FGLS). Table 4, column (f), reports the FGLS estimates. The estimated coefficients were quite similar to those in the baseline. The major differences were that the coefficient of GDP per capita was larger, while the coefficient of the interaction between the measure of compliance with the Directive and piracy rate was more negative. Overall, the estimated impact of compliance with the Rental Directive on movie production was +2.46% (±1.55%).

Comparing the fixed effects estimate with robust cluster variance matrix and the FGLS estimate, we considered that neither was obviously preferable to the other. In the fixed effects estimate (Table 4, column (e)), the coefficients of the real interest rate and piracy were precisely estimated. In the FGLS estimate, the coefficients of the measure of compliance with the Directive and the interaction between the compliance indicator and piracy were precisely estimated. As the results from the two approaches were very similar, in the following robustness checks, we report only the results from the fixed effects estimator.

6. Robustness

[12] For unbalanced panels, this test uses only the observations available for all cross-sectional units. Thus to conduct the test, we had to exclude France, which resulted in 180 observations. The estimation on the smaller samples provided even more compelling results on the impact of the Rental Directive: +4.66% (±2.19%) as compared with +3.97% (±2.09%) from Table 4, column (d).

To check the robustness of the results in Table 4, we also did the following. First, we checked whether our results were driven by a possible outlier country. We estimated the baseline specification excluding one country at a time. Figure 3 depicts the results in terms of the estimated impact of compliance with the Rental Directive on movie production and the corresponding [..]% confidence interval. Evidently, the results were most sensitive to the exclusion of Belgium and the Netherlands. Even so, the estimated impact of the Rental Directive was significantly differently from zero, albeit marginally so.

-- Figure 3: Impact of Rental Directive on movie production: Outlier check --

Next, we checked the sensitivity of our results to the measure of compliance. An alternative indicator is simply the sum of the indicators of compliance with Articles 1 (rental), 1 (lending), 3 (presumption of transfer), 4 (unwaivable right of remuneration), 7 (reproduction), and 9 (distribution).

Table 5, column (a), reports the results. Relative to Table 4, column (e), the major difference is that the coefficient of the interaction between the measure of compliance with the Rental Directive (now the sum of six indicators) and piracy rate was more than halved. Since higher piracy had a smaller effect on the response to the Directive, the estimated impact of compliance with the Directive on movie production was larger, specifically +4.43% (±2.76%).

Another possible source of error in measurement concerned piracy. We used the rate of music CD piracy, which might not perfectly reflect enforcement against movie piracy. Moreover, the data on music CD piracy was only available up to the year [2000]. To check the sensitivity of our results to the measure of piracy, we re-estimated the baseline model using the rate of business software piracy as reported by the Business Software Alliance in place of music piracy.

Table 5, column (b), reports the results. Relative to Table 4, column (e), the major differences were that the overall fit was much worse (R^2 of 0.175 as compared with 0.588) and the coefficient of piracy was positive, albeit not statistically significant. More importantly, from the viewpoint of our policy question, the result was even stronger. The estimated impact of compliance with the Directive on movie production was +8.29% (±3.87%).

Next, we considered the effect of lags between the commissioning and release of movies. In U.S. movie industry, the time from conception to production of print is at least

18 months (Vogel (2004) pp. 53-55). With regard to our context, the Rental Directive was issued in mid-November 1992, which was 17 months before the required changes in law. To the extent of this advance notice, there would not have been any lagged effect on movie production.

Nevertheless, we estimated the baseline specification with all independent variables lagged by one year. Table 5, column (c), reports the results. Relative to Table 4, column (e), the fit was much worse (R^2 of 0.0153 as compared with 0.588). While the coefficients were not very different, they were much less precisely estimated. Even so, the estimated impact of compliance with the Directive on movie production was quite similar, specifically +3.86% (±2.84%).

Besides measurement error and lags, another possible source of bias was omission of relevant explanatory variables. The obvious possibly omitted variable was other legal changes that took effect at the same time as the Rental Directive. Besides the Rental Directive, there were just two major developments in copyright law applicable to the European movie industry in the 1990s (Helberger 2000). They were

- European Copyright Term Directive, which extended the term of copyright to essentially the author's life plus 70 years with effect from July 1995,[13] and
- The WIPO Copyright Treaty, 1996, which created the rights of distribution, rental, and communication to the public.

The WIPO Copyright Treaty was agreed to come into effect three months after thirty member states had deposited instruments of ratification or accession. The Treaty came into effect only in March 2002, following the accession by Gabon. It is unlikely that the WIPO Copyright Treaty would have affected movie production in the 1990s. Moreover, the Treaty substantially overlapped with the Rental Directive.

By contrast, the European Copyright Term Directive took effect around the same time as the Rental Directive. On theoretical grounds, the impact of the Copyright Term Directive on production of creative work is thought to have been minimal since the extension was so far into the future (Akerlof et al. 2002). However, Png and Wang (2007) found …. Accordingly, we estimated the baseline specification including an indicator of compliance with the Copyright Term Directive as an additional explanatory variable.

[13] Directive 93/98/CEE, *O.J.* No. L 290 of 24 November 1993. There was no extension in Germany as its copyright term was already author's life plus 70 years.

Table 5, column (d), reports the results. The estimated coefficients were quite similar to those in the baseline estimate (Table 4, column (e)). The coefficient of the indicator of compliance with the Copyright Term Directive was positive but not significant. The estimated impact of compliance with the Rental Directive on movie production was +4.08% (±2.51%), which was quite similar to and more precise than that with the baseline estimate.

Besides contemporaneous legal changes, another possible omitted variable was government funding. The EU and member states systematically targeted movie production with government funding and tax incentives (Lange and Westcott 2004). However, the only source of data on government incentives for movie production that we could find was the European Audiovisual Observatory's KORDA online database and earlier publications. This provides only information about government funding, and the coverage for the early 1990s is fragmentary. Using the Observatory data, we estimated the baseline specification including government funding as an additional explanatory variable.

Table 5, column (e), reports the results. Owing to the limitations of the data, the number of observations was reduced to 142. The estimated coefficients were quite similar to those in the baseline estimate (Table 4, column (e)). Relative to Table 4, column (e), the fit was much worse (R^2 of 0.272 as compared with 0.588). Strangely, the coefficient of piracy was positive and significant, and the coefficient of government funding was negative, and almost significant. The coefficients of the measure of compliance with the Rental Directive and of the interaction between the measure of compliance and the piracy rate were larger in magnitude and more precisely estimated. However, the estimated impact of compliance with the Directive on movie production was +2.64% (±4.67%).

The poor fit and the counter-intuitive estimates of the coefficients of piracy and government funding were possibly the results of the incomplete data on government funding. Consequently, we are skeptical of the estimate including government funding, as reported in Table 5, column (e).

Yet another possible missing variable is that the Rental Directive took effect together with multiple changes in laws and regulations that improved the overall investment climate across the entire economy. Specifically, the European Union harmonized copyright laws as part of its single-market initiative, and Central and East European countries revised their copyright laws in anticipation of joining the European Union. Hence, any increase in movie production might be due to market expansion and

removal of barriers to intra-European trade rather than the Rental Directive.

The most relevant data that we could find was from the European Audiovisual Observatory, which reports, for each European country, the numbers of movies exhibited in cinemas that were produced domestically and in other European and the United States. In Figure 4, we depict the number of movies from other European countries. There is no obvious upward trend, which suggests that any increase in movie production was not due to the single European market. This inference was supported by regressions reported in Table 5. In column (a), we report estimates of regressions of the numbers of movies produced in other European countries on a time trend. For no country was the time trend positive and significant. In column (b), we report a regression on country and year indicators. None of the year indicators was positive and significant.

-- Figure 4: Numbers of exhibited movies produced in other EU countries --

7. Price Discrimination

So far, we have not considered the mechanics of how the European Rental Directive stimulated movie production. Presumably, the Rental Directive raised the profits that movie studios expected from making movies. Videos provide [..]% of movie studio revenues. For studios, a major impact of the Directive was to allow them to directly discriminate in the sale of video tapes to retailers between those for sale vis-à-vis rental to the end-consumer.

Video retailers both sold and rented tapes. Prior to the Rental Directive, the copyright laws of the various European countries differed in whether movie studios could control rentals of video tapes by retailers. Indeed, it was Warner's attempt to enforce such controls that triggered the issuance of the Rental Directive. Following revision of national laws to comply with the Rental Directive, direct discrimination between tapes for sale vis-à-vis rental became feasible throughout the EU.

To model the impact of the Rental Directive on retail pricing, we follow Varian (2000) and Mortimer (2007). Suppose that direct discrimination is legal. Consider a monopoly movie studio that produces tapes at a marginal cost of c, and sells the tapes for sell-through and rental at wholesale prices of w_s and w_r respectively. For simplicity, suppose that the video retail industry is perfectly competitive, operates with zero marginal cost and zero mark-up on wholesale price, and turns over each rental tape τ times. Since retailers operate with zero mark-up, the retail prices for sale and rental would be of w_s and

w_r/τ respectively. Let the retail demands for sale and rental tapes be $Q_s(w_s)$ and $Q_s(w_s/\tau)$.

Suppose the studio sets wholesale prices to maximize profit

$$\pi = [w_s - c]Q(w_s) + [w_r - c]Q\left(\frac{w_r}{\tau}\right). \quad (2)$$

The first-order conditions are

$$\frac{d\pi}{dw_s} = [w_s - c]\frac{d}{dw_s}Q(w_s) + Q(w_s), \quad (3)$$

and

$$\frac{d\pi}{dw_r} = [w_r - c]\frac{d}{dw_s}Q\left(\frac{w_r}{\tau}\right) + Q\left(\frac{w_r}{\tau}\right). \quad (4)$$

By contrast, if direct discrimination were not legal, we suppose that the studio would maximize (2) subject to the constraint, $w_s = w_r = w$. Generally, in the unconstrained scenario, (2)-(4), the studio would earn higher profit and set different prices $w_s \neq w_r$, while in the constrained scenario, the studio would earn lower profit and be required to set a single price, *w*. Accordingly, we expect that compliance with the Rental Directive would be associated with discrete changes in the retail prices and rental rates.

The analysis does not provide any unequivocal prediction as to the direction of change of the prices and rental rates. To the extent that the demand for rental tapes is more price-elastic than the demand for sale tapes, compliance with the Rental Directive would lead to an increase in rental rates and a reduction in sale prices.[14] However, in the case of the United States, Mortimer (2007) calculated that, for a sample of high-grossing movies released in 2000-01, direct discrimination would have resulted in rental rates that would have been 12% or 19% *lower* than without direct discrimination.[15]

[14] Here, it is important to emphasize the *ceteris paribus* assumption, in particular, that the compliance with the Rental Directive was not contemporaneous with introduction of revenue-sharing contracts to resolve double marginalization between studios and retailers (Varian 2000).

[15] Under the first sale doctrine of U.S. copyright law, direct discrimination between tapes for sale vis-à-vis rental is not legal. Mortimer (2007) estimated the retail demand for video-tape sales and rentals. She then calculated the average price and rental rate that would have maximized profit if direct discrimination were legal. For tapes initially priced for rental, direct discrimination would have reduced the rental rate by 12% from $2.84 to $2.51, while for tapes initially priced for sell-through, direct discrimination would have reduced the rental by 19% from $3.04 to $2.45. The implication of direct discrimination on the retail price of video-tapes was ambiguous because,

We collected data on retail prices and rental rates of videos for the same 17 European countries over the period 1993-2005 from Euromonitor's Global Market Information Database (GMID). We should immediately caution that the retail prices and rental rates pertain to all videos, including those originating from the U.S. and other countries. Unfortunately, the available information did not distinguish the country of origin. Hence, our empirical study of pricing would not exactly align with our study of European movie production reported in Sections 4-6 above.

The GMID pricing data was very fragmentary for the early years, and indeed data for 1995 were almost entirely missing. We converted the prices and rental rates at the prevailing exchange rates to European Currency Units (Euros). Table 7 provides summary statistics.

-- Table 7: Videos: Descriptive statistics –

To provide an overview of the impact of the Rental Directive, Figure 5 illustrates, for each country, the prices and rental rates of video tapes and the degree of compliance with the Directive over the period 1993-2005. The graphs suggest a …. correlated with the increase in compliance with the Rental Directive.

-- Figure 5: Video pricing and compliance with Rental Directive –

To further investigate the impact of the Rental Directive on video pricing, we regressed the natural logarithms of video rental rates, rental volumes, sale prices, and sales volumes on various national characteristics that might possibly affect the demand for video sales and rentals – population, GDP per capita, computer ownership and internet access, piracy, and the measure of compliance with the Rental Directive.

Generally, in all of these estimates, we encountered serial correlation and heteroscedasticity. We addressed these with two alternative estimators – fixed effects with robust cluster covariance matrix and feasible GLS. We preferred the feasible GLS estimators as they tended to provide more precise estimates of the impact of the Rental Directive.

Table 8, columns (a)-(c), reports results for video rental rates. Compliance was associated with a 3.86%(±0.95%) increase in rental rates. Table 8, columns (d)-(f), reports

---

under the strategy of pricing initially for rental, the retail price would be first set high and then, when the rental market had been saturated, the retail price would be reduced.

the corresponding results for the volume of video rentals. Compliance with the Rental was associated with the volume of video rentals being 2.91%(±1.87%) lower. The reduction in rental volume was economically consistent with higher rental rates. Taking account of both the increase in rental rate and the decline in rental volume, the Rental Directive was associated with an average increase in revenue of about [3.86% – 2.91%] = 0.85%. Using the mean rental rate and volume from Table 7, this amounted to 0.85% x 2.88 x 29.1 = €800,000 per country annually.[16]

-- Table 8: Video rentals --

Table 9 reports results for selling prices of video tapes and the corresponding sales volumes. Compliance with the Rental Directive was associated with a 1.23%(±1.21%) reduction in selling price and a 3.04%(±1.88%) increase in sales volume. The increase in sales volume was economically consistent with lower prices.[17] Taking account of both the reduction in selling prices and the increase in volumes, the Rental Directive was associated with an average increase in revenue of about [3.04% – 1.23%] = 1.81%. Using the mean rental rate and volume from Table 7, this amounted to or 1.81% x 11.9 x 11.8 = €2.541 million per country annually.[18]

-- Table 9: Video sales --

8. Concluding Remarks

We investigated the impact of the European Rental Directive on the production of movies in 17 countries during the period 1993-2005. We found that, on average, revision of national laws to comply with the Directive was associated with an increase in movie production ranging between 2.46% (±1.55%) and 3.97% (±2.85%). Importantly, the increase in production was higher in countries with lower rates of piracy.

[16] The corresponding increase in revenue using the fixed-effects estimates was [5.35% – 5.10%] x 2.88 x 29.1 = €210,000 per country annually.

[17] Our model of video pricing, (2)-(4), did not consider that, if movie studios could not directly price discriminate, they might discriminate indirectly by pricing tapes high, targeting the rental demand, and then, after some months, cutting the price low for sell-through (Mortimer 2007). The Rental Directive would allow studios to set two prices – one for rental and another for sell-through. Hence, the impact on the selling price would be two-fold: lower to the extent that the price need not be balanced against the price for rental tapes, but possibly higher to the extent that studios would no longer cut prices for sell-through after some months.

[18] The corresponding increase in revenue using the fixed-effects estimates was [2.00% – 1.76%] x 11.9 x 11.8 = €337,000 per country annually.

These findings were robust to exclusion of a possible outlier country, and various specifications, including an alternative measure of compliance with the Rental Directive and an alternative measure of piracy, other contemporaneous changes in copyright law, and changes in government funding.

These findings were bolstered by a study of the impact of the Directive on video sales. We found that on average, revision of national laws to comply with the Directive was associated with a 0.85% increase in revenue from video rentals and a 1.81% increase in revenue from video sales.

These results are significant as there have been very few systematic empirical analyses to show any impact from changes in copyright law on the production of creative work (Png 2006). The most obvious direction is to study the production of creative work more deeply, to better understand the intermediate links between copyright law and creative output. How does copyright law affect investment in creative activity on two margins – the number of titles and the investment in each title? And, how do these investments translate into the quantity and quality of creative output such as movies, books, and recorded music?

The other direction for future work is to measure the impact of copyright law on the use of existing creative work, and specifically, on the benefit to end-users as well as investment in creations that build upon earlier work.

With the results from these studies, it would then be possible to gauge the fundamental trade-off in copyright law between the incentive to create new work and the loss from restricting use of existing work. However, the key challenge in all of these directions for future work is to acquire the relevant data.

References

Akerlof, George A., et al., "The Copyright Term Extension Act of 1998: An Economic Analysis", Washington DC: AEI-Brookings Joint Center for Regulatory Studies, 2002.

Allen Consulting Group, "Copyright Term Extension: Australian Benefits and Costs", Report Commissioned by the Motion Picture Association, July 2003. www.allenconsult.com.au/resources/MPA_Draft_final.pdf [Accessed, August 23, 2005].

Baker, Matthew J., and Brendan M. Cunningham, "Law and Innovation in Copyright Industries", U.S. Naval Academy, November 2005.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-In-Differences Estimations?" *Quarterly Journal of Economics*, Vol. 119 No. 1, February 2004, 249-275.

Boldrin, Michele, and David K. Levine, "The Case Against Intellectual Property", *American Economic Review*, *Papers and Proceedings*, Vol. 92 No. 2, May 2002, 209-212.

Donald, Stephen G., and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data", *Review of Economics and Statistics*, Vol. 89 No. 2, May 2007, 221-233.

European Commission (EC), Transposition des directives sur le droit d'auteur et les droits voisins dans la législation des Etats members, Contrat d'étude ETD/99/B5-3000/E/15, undated.

Gallini, Nancy, and Suzanne Scotchmer, "Intellectual Property: When is it the best incentive system?" in Adam Jaffe, Joshua Lerner and Scott Stern, eds, *Innovation Policy and the Economy*, Vol. 2, Cambridge: MIT Press, 2002, 51-78.

Geller, Paul Edward, ed., *International Copyright Law and Practice*, 2 vols. New York, NY: Matthew Bender & Company, October 1999.

Helberger, Natali, "Copyright and Related Rights in the Audiovisual Sector", in *IRIS Focus: Copyright Law in the Digital Age*, European Audiovisual Observatory, Strasbourg, France, 2000.

Hui, Kai-Lung, and Ivan Png, "On the Supply of Creative Work: Evidence from the Movies", *American Economic Review*, *Papers and Proceedings*, Vol. 92 No. 2, May 2002, 217-220.

IFPI (International Federation of the Phonographic Industry), *The Recording Industry: Commercial Piracy Report, 2004*, London.

Khan, B. Zorina, "Does Copyright Piracy Pay? The Effects of U.S. International Copyright Laws on the Market for Books, 1790-1920", National Bureau of Economic Research, Working Paper 10271, January 2004.

Landes, William M., and Richard A. Posner, "An Economic Analysis of Copyright Law", *Journal of Legal Studies*, Vol. 18, June 1989, 325-363.

Landes, William M., and Richard A. Posner, *The Economic Structure of Intellectual Property Law*, Cambridge, MA: Belknap Press, 2003.

Lange, Andre, and Tim Westcott, *Public funding for film and audiovisual works in Europe – A comparative approach*, European Audiovisual Observatory, Strasbourg, 2004.

Lessig, Lawrence, *The Future of Ideas: The Fate of the Commons in a Connected World*, New York: Random House, 2001.

Liebowitz, Stan J., "Copying and Indirect Appropriability: Photocopying of Journals", *Journal of*

Political Economy, Vol. 93 No. 5, October 1985, 945-957.

Motion Picture Association of America, http://www.mpaa.org/piracy_WhoPiracyHurts.asp [Accessed, August 4, 2006].

Mortimer, Julie Holland, "Price Discrimination, Copyright Law and Technological Innovation: Evidence from the Introduction of DVDs", *Quarterly Journal of Economics*, Vol. 122 No. 3, August 2007, 1307-1350.

Plant, Arnold, "The Economic Aspects of Copyright", *Economica*, Vol. 1 No. 2, May 1934, 167-195.

Png, I.P.L., "Copyright: A Plea for Empirical Research", *Review of Economic Research on Copyright Issues*, Vol. 3 No. 2, 2006, 3-13.

Png, I.P.L. and Qiu-hong Wang, "Copyright Duration and the Supply of Creative Work: Evidence from the Movies", Dept of Information Systems, National University of Singapore.

Quah, Danny, "Almost efficient innovation by pricing ideas", Dept of Economics, London School of Economics, June 2002.

Reinbothe, Jorg, and Silke von Lewinski, *The E.C. Directive on Rental and Lending Rights and on Piracy*, London: Sweet & Maxwell, 1993.

Reynolds, Taylor F., "Quantifying the Evolution of Copyright and Trademark Law", PhD dissertation, American University, Washington, D.C. 20012, 2003.

Varian, Hal, "Buying, Sharing and Renting Information Goods", *Journal of Industrial Economics*, Vol. 48 No. 4, December 2000, 473–488.

Vogel, Harold L., *Entertainment Industry Economics*, Cambridge, UK: Cambridge University Press, 6th Edition, 2004.

Waterman, David, with Weiting Lu, Appendix K: "Movie Genre Analysis", in David Waterman, *Hollywood's Road to Riches*, Cambridge, MA: Harvard University Press, 2005.

Watt, Richard, "The Past and the Future of the Economics of Copyright", *Review of Economic Research on Copyright Issues*, Vol. 1 No. 1, June 2004, 151-171.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity", *Econometrica*, Vol. 48 No. 4 (May 1980), 817-838.

Wooldridge, Jeffrey M., *Introductory Econometrics: A Modern Approach*, Mason, OH: Thomson, 2006.

Waiting to copy: the dynamics of the market for technology

Emeric Henry*and Carlos J. Ponce†

MARCH 2008

Abstract

We examine the profits of an innovator in the absence of intellectual property rights protection but in the presence of a market for technology. Potential imitators can enter the product market by either copying the invention at a cost or by obtaining a license from the inventor. As imitators enter the product market, they also compete with the innovator on the market for technology. The price of the technology transfer licenses falls as competition to provide them increases. This creates an incentive for imitators to wait to enter the market, in the hope that a competitor will enter before them. Furthermore, we show that if there is free entry of imitators, competition on the market for technology can be so fierce that it dissuades any imitator to incur the initial entry cost, thus guaranteeing monopoly profits for the inventor. Our results challenge the traditional view that in the absence of intellectual property protection, the innovators' rents will be competed away.

Jel Codes: L24, O31, O34
Keywords: Innovation, Licensing, Intellectual Property Rights

1 Introduction

*Economics Subject Area, London Business School (email: ehenry@london.edu)
†Department of Economics, Universidad Carlos III de Madrid (email: cjponce@eco.uc3m.es)

The size of "markets for technology"[1] has been steadily growing over the past years. Although data is imprecise, Degnam (1998) estimates that US firms, individuals and governmental organizations received in 1998 in the order of $73 billion in licensing revenues.[2] Licensing is essential for overall innovation. It favors for instance the dissemination of ideas. It also allows inventors to reap the benefits of their innovation without needing to invest in production. This is indeed the major explanation for licensing behavior provided in the literature. However, Arora et al. (2002) note that large established companies, who produce on the market, such as IBM or Union Carbide in the chemical industry, are very active licensors. They also present partial evidence based on data from the chemical industry, demonstrating that licensing is more widespread in product groups that are more homogenous. This evidence suggests that producers also license to their competitors.

In this paper we show that this type of competitive licensing behavior can have major implications for the profits of innovators. We demonstrate that the dynamics on the market for technology can protect the rents of the innovator even in the absence of intellectual property rights and lead to endogenous delay of entry by potential imitators. Furthermore, an increase in the number of potential entrants can paradoxically have a positive effect on the innovator's rents. In the case of free entry of imitators, the innovator can retain full monopoly profits. These results are a strong challenge to the traditional view on intellectual property rights protection.

We consider a model where an innovator has developed a new technology. She faces two potential imitators who need to incur a cost κ to imitate the innovation. An alternative to entry by imitation is to purchase a license from the innovator who can transfer the idea at a cost ϵ smaller than the imitation cost. At each period the innovator and the firms that entered compete on the product market using the technology. In the absence of intellectual property rights and if the market for technology does not exist, both imitators will immediately enter the product market by copying the invention, provided the imitation cost is low enough. The rents of the innovator will immediately be competed away and this could discourage her from initially investing in research. This represents the traditional justification for intellectual property rights protection. We demonstrate in this article that this logic is no longer valid when the possibility of licensing is considered.

When a market for technology exists, once an imitator has entered the product market he also becomes a competitor of the innovator on the market for technology. The innovator and the first imitator compete on prices, dissipate their licensing profits and provide the license at a minimal price of ϵ.[3] The price of the license thus dramatically falls after entry of the first imitator, leading to

[1] We employ the term used in Arora et al (2002): technology refers to knowledge rooted in engineering and science, but also drawn from production experience. It covers both tangible knowledge (designs, formulae...) as well as know-how.

[2] An analogous figure is reported in Arora et al (2002)

[3] This is the case in one particular equilibrium of the subgame following entry and we provide condition under which this is indeed the unique equilibrium. The condition imposes that profits do not decline too quickly with the number of competitors on the market.

higher profits for the second imitator. The dynamics of competition on the market for technology thus naturally leads to a war of attrition where both imitators delay entry in the hope that their competitor will enter before them.

The expected delay in entry and equilibrium profits of the innovator increase with the imitation cost and decrease with the transfer cost. Furthermore, as κ converges to triopoly rents, we show that the profits of the innovator converge to monopoly profits, in other words the profits he would obtain if intellectual property rights were protected. We illustrate those dynamics in a series of simulations that reveal the non-montonicity of licensing profits when the imitation cost increases. An increase in κ has two effects: it increases the instantaneous value of licensing profits but also delays entry and therefore delays the date at which these profits are obtained. Whereas the first effect could dominate for small values of κ the second will always dominate for larger values.

The first part of the paper assumes that the innovator faces only two potential imitators. It is natural to examine whether the conclusions are altered when we consider free entry of imitators. We show that such a situation reinforces our argument. Under free entry, there always exists an equilibrium where the innovator retains full monopoly rents. This equilibrium is based on extreme competition on the market for technology after the initial entry of the first imitator. The potential entrants given this competition expect future profits to be insufficient to cover the initial entry cost. They shy away from entry and leave the innovator with monopoly rents. More importantly we derive a condition under which this is the unique equilibrium, condition that is generally satisfied for linear demand and Cournot competition. This condition requires that the rate at which profits decrease with the number of competitors on the product market is not too high. Paradoxically, a not very competitive product market will create a very competitive licensing market.

Finally we examine a number of extensions of the model. We first determine if allowing more complex contracts involving fixed fees and royalty payments will alter our conclusions. We show that royalty rates will not be used in equilibrium and therefore the conclusions on endogenous delay remain unaffected. This interesting result represents a general contribution to the literature on licensing contracts. In a duopoly case, it is well known that royalty rates can increase the joint surplus of the licensor and licensee by restricting the quantity produced. Our results suggests that when a bilateral contract is signed in a larger oligopoly, a pure fixed fee contract will be used. Indeed, if a royalty rate is included to decrease the production of the licensee, this will benefit both the licensor and her competitor, who might react by increasing his own production. We show in a particular case that the second effect will dominate and including a positive royalty rate will tend to decrease the bilateral surplus.[4]

[4] As the competitor increases his production, this further decreases the profits of the licensee and the licensor is unable to fully compensate using the fixed fee as she could in the duopoly case.

Our paper challenges the view that in the absence of intellectual property rights, the rents of innovators will be competed away. To reach this conclusion, we took into account the dynamics of the market for technology, a factor ignored in the existing literature. It has been pointed out, in contributions that we will examine in section 6, that intellectual property can facilitate licensing in the presence of information asymmetries between the innovator and the imitators regarding the value of the innovation. These asymmetries can make potential licensees wary to sign a contract. On the other hand, in the absence of intellectual property rights, if the innovator reveals information to convince the licensees, she epxoses herself to the risk that they will copy the innovation without making any type of payment. We believe that information asymmetries are small in the context we are studying and we ignore them in our model. In our environment the value of the innovation is publicly observed: the innovator already has a completed version of the technology and is already implementing it. However, we do not refute the potential importance of some type of protection for earlier stage research of unproven commercial potential. We challenge the more basic justification of intellectual property rights that states that the innovator's rents will be immediately competed away.

We examine fully the links with the existing literature, including papers on licensing under information asymmetries in section 6 after having presented our results. In section 2 we present the details of the model. In section 3 we present our main result on endogenous delay and innovator's rents in the case of two potential imitators. In section 4 we show that these conclusions hold more generally under free entry of potential imitators. Surprisingly, a more competitive environment reinforces our argument. Finally in Section 5, we discuss some extensions of the model and in particular the inclusion of royalty rates.

2 The Benchmark Model

We consider a multistage game of complete information. Time is indexed by $t\Delta$ for $t \in T := \{0, 1, 2, ...\}$, where $\Delta \in \mathbb{R}_{++}$ is the *real* length of each time period.[5] Three players interact at each time period: an inventor ('she'), denoted by s, and a set of two individuals, $\mathcal{I} := \{a, b\}$, each of which is called an imitator ('he'). The inventor before the start of the game has developed an invention which represents an improvement over the previous state of the art.[6]

Each imitator $i \in \mathcal{I}$ may use at each date $t \in T$ an *imitation* technology to obtain a *perfect* version of the innovation. More formally, by paying an imitation fee $\kappa \in \mathbb{R}_{++}$ at time $t \in T$, an imitator implements at date $t \in T$ (i.e., *instantaneously*) a perfect version of the invention. The imitation fee

[5] From now on, to avoid cumbersome notation, unless it is absolutely necessary, we will denote a time period by $(t + z)$ instead of $(t + z)\,\Delta$, $\forall t, z \in T$.

[6] The reader may think of the innovation as being either a product improvement or a cost reducing innovation.

can be interpreted as one-time sunk cost that must be incurred by an imitator to enter the market by imitation.

An alternative to imitation is to sign a contract with the innovator who will transfer the technology and possibly 'know-how'. The innovator being the creator of the innovation, possesses technological abilities which, if they were transferred to an imitator, would allow him to save his imitation costs. This transfer is however costly for the innovator. These costs include the general costs of contracting (bargaining, enforcing the agreement), but also the costs of transferring intangible know-how.[7] We denote by $\epsilon \in \mathbb{R}_{++}$ the cost of transfer for the inventor. Throughout this paper, we make the assumption that the transfer cost is smaller than the imitation cost $\epsilon < \kappa$. The trades between the inventor and imitators for licenses are part of what is called the *market for technology*. We describe later on the dynamics in this market.

We now describe competition in the *product market*. First we note that if an imitator has signed a license with the innovator or copied the invention at t, he can immediately start producing in the current period. The profits earned by each agent, in general, depend upon demand conditions, the market game being played and the features of the innovation. For our purposes it is sufficient to specify equilibrium profits in reduced form. Also, for simplicity, we normalize the profit functions of the imitators in such a way that their profits when they do *not* use the innovation are equal to zero.[8] Besides, we suppose that when an imitator implements the innovation, he will be in the same position as the innovator. In other words, we assume that the innovator and each *active* imitator (i.e., an imitator who has implemented the innovation) obtain the *same* profit rate regardless of the mode of entry (imitation or contracting). More formally, the profit rate (gross of the imitation fee or payments between the agents) satisfies: $\pi_1 > \pi_2 \geq \pi_3 \geq 0$, where π_j denotes the profit rate when j firms compete on the market.[9] Finally, all agents are risk neutral and maximize their expected discounted profits. They have a common discount rate $r \in \mathbb{R}_{++}$ and the discount factor associated with a lag of real time $\Delta \in \mathbb{R}_{++}$ is $\delta(\Delta) := \delta = \exp(-r\Delta) \in (0,1)$.[10]

2.1 Contracting

We suppose that the innovator makes a take it or leave it offer to the potential imitators where the contract involves only a fixed fee. We explore some alternative modelling choices in section 5. We first examine the case of contracts involving both fixed fee and royalty rate. We then study the introduction

[7] See Teece (1977)

[8] This is just a convenient analytical normalization. Our model is sufficiently general to encompass situations in which the innovation is drastic and non-drastic.

[9] Observe that the profit *rate* is the profit that an agent obtains at each instant of real time.

[10] The product market game and our economy is 'stationary' in the sense that profits depend only on the number of active agents in the product market and not of the time period.

of more elaborate bargaining games between innovator and imitators. Section 5 demonstrates that these changes do not fundamentally impact our results.

As we pointed out in the introduction we assume that there are no information asymmetries between the innovator and the potential imitators. In our context, the value of the innovation is publicly observed as the innovator is already implementing it. We therefore assume that there are no hurdles to contracting apart from the traditional contracting costs included in the transfer cost ϵ.

2.2 Competition in the Market for technology

An imitator can implement the innovation by either (i) using the imitation technology or by (ii) contracting with the innovator. To clarify the terminology we will use throughout this paper, when an imitator implements the innovation we will describe him as entering the product market. The cost of imitation or the cost of the contract can thus be viewed as a sunk cost of entry. In this section we clarify whether the mode of entry of the imitator will determine whether he can sell licenses in future periods, in other words whether he will compete on the market for technology.

When the imitator signs a contract with the innovator it is natural to assume he will become not only her *future* competitor in the product market but he will also be her *future rival* in the market for technology.[11] The knowledge and ideas transferred by the inventor will now be a component of the human capital of the imitator who acquired these services and he will be able to transfer them for a fee to future imitators. We assume that the transfer cost will be the same for the imitator as for the initial inventor (equal to ε).

The question of whether the imitator who entered by copying will also become a competitor on the market for technology is a more delicate issue. When he imitates, does he also learn the know-how that is possessed by the innovator? The answer depends on what exactly an imitation technology is, an issue largely unexplored in the literature. We could face two extreme cases. In the first the mode of entry has no impact on future competition: the imitator who copied the innovation will compete on the market for technology. In the other extreme case, the imitator only gets a copy of the innovation and does not have the knowledge necessary to transfer the technology. We suppose in this paper that imitators will become competitors on the market for technology *only if* they contracted with the innovator. We will point out that our argument would also hold if we had considered the other extreme case.

2.3 Timing of the game

We can now present the timing of this multi-stage game of complete information.

[11] This supposes that the innovator does not include a clause in the contract that prevents the licensee to provide licenses in future periods. We examine this issue in the main text.

At a history of the game where the inventor faces no competitor on the market for technology, the timing at period $t \in T$ is the following:

(i) The innovator announces the contracts she offers

(ii) Then, imitators after observing the collection of contracts offered by the innovator decide simultaneously whether to implement the innovation or not and choose their mode of entry (copying or contracting)

(iii) The innovator and all the imitators that implemented the innovation at t or at previous periods, compete on the product market.

At a history of the game where an imitator (that we will call the leader) has previously entered by contracting with the innovator, the timing at period $t \in T$ is the following:

(i) The innovator and the leader imitator announce the contracts they offer

(ii) Then, the follower imitator, after observing the contracts decides whether to implement the innovation or not and chooses his mode of entry

(iii) The innovator and all the imitators that implemented the innovation at t or at previous periods, compete on the product market.

All players observe the complete history of the game. A history of the game in period t is a sequence of contracts proposed by the innovator (and possibly by one of the imitators), a sequence of implementation decisions taken by the imitators and a sequence of decisions of *how* to implement the innovation. Let $\mathcal{H}^t$ denote the set of all possible histories in period t. A pure strategy for the innovator is a sequence of maps $\left\{\sigma_s^t\right\}_{t\in T}$ where σ_s^t maps $\mathcal{H}^t$ into collection of contracts. A pure strategy in period t, σ_i^t, for imitator $i \in \mathcal{I}$ who has not yet implemented the innovation, is a map from histories (including the collection of contracts offered by the innovator in period t) to an implementation decision and to a decision of how to implement the innovation. A pure strategy in period t, σ_i^t, for imitator $i \in \mathcal{I}$ who has previously entered by contracting with the innovator, is a map from histories into a collection of contracts. Hence, a pure strategy for imitator i is a sequence of maps $\left\{\sigma_i^t\right\}_{t\in T}$. We will be interested in behavior strategies of the imitators over their implementation decisions. For that, let $(\psi_i)_{i\in\mathcal{I}}$ denote the (stationary) probability of implementing the innovation at date t.[12] We concentrate on subgame-perfect equilibria.

3 Main Results

3.1 Benchmark case: No contracting

We start by computing the equilibrium in the benchmark case where the market for technology is considered inexistent and contracting is therefore unfeasible. In this case, entry may only occur through

[12] Extending the above definitions to include behavior strategies is an obvious exercise.

imitation.

It is important to note that although we consider a world where intellectual property rights are not protected, the imitation cost works as an entry barrier determining a natural measure of protection for the inventor. Consider, for example, the situation in which $\kappa > \int_0^\infty \pi_2 \exp(-rt)\, dt = \frac{\pi_2}{r} := \Pi_2$. Because the imitation fee is strictly higher than the present value of duopoly profits, imitation will never occur in equilibrium. The innovator will therefore retain monopoly profits $\Pi_1$even though intellectual property rights are not protected.[13] To make our problem interesting we therefore make the following assumption on the imitation cost:

ASSUMPTION 1: $\kappa < \Pi_3$

Assumption 1 guarantees that the market can accommodate both imitators and that the dynamics in the market for technology are interesting.

PROPOSITION 1: *If the market for technology does not exist, under Assumption 1:*
(i) There is a unique subgame perfect equilibrium in which both imitators imitate at time $t = 0$
(ii) The profits of the innovator and the imitators are equal to Π_3 *and* $\Pi_3 - \kappa$ *respectively.*

PROOF. *See Appendix.*

In the case where contracting with the innovator is not feasible, it is a dominant strategy for both imitators to enter immediately. Indeed, there is no benefit from delaying imitation since the entry cost will remain fixed throughout the game at the imitation cost κ. Furthermore, by delaying entry, the imitators will loose profits in the current period. Therefore, if entry occurs it will take place at $t = 0$ with probability one. Assumption 1 guarantees that entry is indeed profitable for both imitators.[14]

Proposition 1 summarizes the conventional wisdom justifying intellectual property rights protection. In the absence of IP rights, imitators will enter immediately following a successful innovation and will compete away the rents of the innovator. Foreseeing this risk, innovators might thus shy away from initially investing in research. The purpose of our paper is to challenge this line of thought and to show that delay can actually occur in equilibrium when a market for technology exists.

3.2 Imitation and Contracting

We now consider the case where a market for technology exists and the innovator can contract with the potential imitators. We first determine the players' equilibrium behavior in the subgame following entry by one single imitator.[15]

[13] From now on we use $\Pi_j := \frac{\pi_j}{r}$ for $j = 1, 2, 3$.
[14] They both obtain profits of $(\Pi_3 - \kappa) > 0$.
[15] More specifically the subgame following a history in which no imitator has entered prior to $t \in T$ and at which at t only one imitator, say $i \in \mathcal{I}$, has decided to implement the innovation.

We initially want to highlight that if imitator i enters at t by contracting with the innovator, there always exists an equilibrium where imitator j immediately follows, enters at $t+1$ by contracting and pays a minimal price of ϵ for the license. This equilibrium rests on extreme competition on the technology market between the innovator and imitator i to provide a license: the equilibrium strategies involve offering the license at a price of ϵ every period.[16] Although this equilibrium seems natural, there are other more collusive equilibria where both the innovator and imitator i keep higher prices for some time period. The following lemma shows that the equilibrium with immediate entry will be the unique equilibrium under the following assumption:

ASSUMPTION 2: $2\Pi_3 - \Pi_2 \geq \epsilon$

The importance of this assumption will become clear when we present the intuition of Lemma 1. The assumption can be viewed as restricting the rate at which profits decrease when the number of competitors increase. It is essential to note that for linear demand and Cournot competition, assumption 2 will always be satisfied for values of ϵ that are small enough. Under assumption 2, we obtain the following result:

LEMMA 1: *Under Assumption 2, in all SPNE, if no imitator has entered prior to $t \in T$ and imitator $i \in I$ has entered at t, then imitator j will enter at period $t+1$. Furthermore, if imitator i entered by contracting with the innovator, imitator j will enter by contracting and pay a price of ϵ for the license.*

PROOF. *See Appendix.*

The intuition for the second part of Lemma 1 is clear: in any equilibrium, at a date where a license is signed, it will be signed at the minimal price of ϵ or one of the competitors would have an incentive to deviate and offer a lower price. The concern however is that there might exist equilibria where the innovator and imitator i initially keep high prices such that the license is not immediately signed. Assumption 2 guarantees that there will be a profitable deviation in such a candidate equilibrium. The intuitive interpretation is that if twice triopoly profits is greater than duopoly competitors have an incentive to offer an extra license at a price close to triopoly profits. We now make this interpretation more precise. Consider an equilibrium where a license is signed at date $\tau > t+1$ for a price of ϵ. If one of the competitors on the market for technology deviates and offers a license at the previous period, the maximum price he can charge is $p_{\max} = (1-\delta)\Pi_3 + \delta\epsilon$ (imitator j would earn triopoly profits for an extra time period and would not pay the license fee the next period). The deviation that consists in offering a license at $\tau - 1$ at a price of $p_{\max}$ is therefore profitable if $(1-\delta)\Pi_3 + \delta\Pi_3 + p_{\max} - \epsilon \geq (1-\delta)\Pi_2 + \delta\Pi_3$.

[16] If her competitor follows such a strategy, the inventor has no incentive to deviate as entry will occur regardless of her contract offer.

Rearranging the terms this condition amounts to assumption 2. Under these conditions, the unique equilibrium will be one where a license is immediately signed. In the rest of this section we will concentrate on this equilibrium

Based on the result of Lemma 1, we can describe the expected profits of an innovator who enters at period t if the competitor has not entered before t (t included). If imitator i enters by copying, his expected profits are:

$$\upsilon_i(C) = \int_0^{\Delta} \pi_2 \exp(-rt)\, dt + \delta\Pi_3 - \kappa = [1-\delta]\,\Pi_2 + \delta\Pi_3 - \kappa$$

He obtains flow profits π_2 this period, and given the result of Lemma 1, expects the competitor to immediately follow and decrease his flow profits to triopoly profits π_3 thereon.

On the other hand, if imitator i enters by contracting with the innovator, his expected profits are

$$\upsilon_i(L) = [1-\delta]\,\Pi_2 + \delta\Pi_3 - p_i^t$$

The only distinction between $\upsilon_i(C)$ and $\upsilon_i(L)$ is the cost of entry. In particular, by contracting with the innovator (action L), imitator i does not expect to obtain licensing revenues in future periods. Indeed, according to Lemma 1, competition on the market for technology will reduce licensing profits to zero. The nature of contracts offered by the innovator will therefore determine which mode of entry is chosen by imitators. This is explored in the following lemma.

LEMMA 2: *In all SPNE, at every date $t \in T$ where no imitator has entered, the imitator offers a contract at a price $p^* = \kappa$.*

PROOF. *See Appendix.*

To understand the intuition of Lemma 2, consider a period $t \in T$ where an imitator i decides to enter. Suppose also that no imitator has entered prior to that date. According to the results of Lemma 1, the mode of entry (copying or contract) will only influence the licensing revenues and not the profits obtained on the market.[17] Furthermore, offering a price of κ for the contract maximizes licensing revenues. Indeed, if the price of the license is set strictly greater than κ, imitator i will enter by copying. At best the innovator can then hope to sign a license in the next period at price κ with the second entrant. On the other hand, the innovator would not want to offer a license at a price less than κ as it would reduce her licensing profits. It is therefore clear that the innovator's strategy that maximizes licensing revenue and thus overall profits is to offer the license at a price of κ. The results

[17] The profits on the market will be π_2 this period and π_3 thereon, independently of the mode of entry.

of Lemma 1 and 2 can now be used to characterize more specifically the subgame perfect equilibria of the game.

PROPOSITION 2: *Under Assumption 2, if contracting is feasible, in any SPE in which the imitators implement the innovation with positive probability:*

(i) imitators will always enter through contracting

(ii) if the imitators do not enter simultaneously, the first imitator to enter pays a price $p^* = \kappa$ *and the second a price* ϵ.

(iii) if the imitators enter simultaneously they both pay a price $p^* = \kappa$.

PROOF. *See Appendix.*

Proposition 2 indicates that copying will never occur on the equilibrium path. It will always be preferable for the innovator to obtain the licensing revenues given that she cannot prevent entry. The second key lesson that we draw is that the price of the license will be dramatically decreased for the second innovator. This influences the expected profits of the players, in a manner made explicit in the following corollary.

COROLLARY 1*: If the first imitator enters at date t, then the expected payoffs at t are:*

- $\upsilon_l = (1-\delta)\,\Pi_2 + \delta\Pi_3 - \kappa$ *for the first imitator*
- $\upsilon_f = \delta\,(\Pi_3 - \varepsilon)$ *for the second imitator*
- $\upsilon_B = \Pi_3 - \kappa$ *if they both enter simultaneously*

The result in Corollary 1 is essential to understand the dynamics of the mechanism we uncover. If the time period is sufficiently small, we see that the game corresponds to a war of attrition. Both imitators have an incentive to wait to enter to obtain the license at a smaller price. We show in the next section that this generates endogenous delays in entry. It however differs from the war of attrition game defined in Fundenberg and Tirole in the sense that the payoff of the follower υ_f is not equal to the payoffs υ_B of the imitators if they both enter simultaneously.

3.3 The Dynamics of Imitation: Delay and Innovative Rents without IP

The results of the previous section will allow us to characterize the equilibrium timing of imitation. We will show that delays will occur in equilibrium and that the innovator will retain monopoly rents for some time even in the absence of intellectual property rights. We determine the symmetric stationary equilibrium of this game in behavioral strategies.

Let us denote by ψ the probability of implementing the innovation at any time period. For a stationary symmetric strategy profile to be an equilibrium, imitators must be indifferent between the

implementation times in the support of their randomization. Therefore a *necessary* condition for an equilibrium in behavioral strategies to exist is that the payoff for implementation at time $t \in T$ (conditional on the innovation not having been implemented yet) be equal to the payoff for waiting one additional time period and implementing the invention at $(t+1) \in T$. The necessary condition can be expressed in the following way

$$V(\psi) := \psi v_B + (1-\psi)v_l \stackrel{\Delta}{=} \psi v_f + (1-\psi)\delta V(\psi) := W(\psi) \tag{1}$$

or equivalently

$$\mu(\psi) = \psi\left[v_f - v_B\right] + (1-\psi)\left[\delta V(\psi) - v_l\right] = 0 \tag{2}$$

where $V(\psi)$ is the *value of implementation* in the current period, $W(\psi)$ is the *value of waiting* an additional time period to implement the innovation and $\mu(\psi) := W(\psi) - V(\psi)$ is the *net* value of waiting. To understand the factors determining the timing of imitation, it is informative to interpret equation (2) characterizing the net value of waiting. With probability ψ the competitor has implemented the innovation in the current period. In that case the net value of waiting is $v_f - v_B$.[18] On the other hand with probability $(1-\psi)$ the competitor does not enter in the current period. In that case, waiting is costly as it sacrifices profits during the current period without changing the continuation value of the game. In an equilibrium in behavioral strategies these two countervailing incentives must balance each other.

The previous discussion highlights the fact that a necessary condition for an equilibrium in behavioral strategies to exist is that $v_f - v_B$ be strictly positive. This is the case if the time period is sufficiently small ($\delta_0 > \frac{\Pi_3 - \varkappa}{\Pi_3 - \varepsilon}$), thus guaranteeing that the benefits from obtaining the license at a smaller price dominates the lost triopoly profits in the current period. We will also see that another condition on the discount rate needs to be imposed to guarantee uniqueness of the symmetric equilibrium. These ideas are formalized in Assumption 3.

ASSUMPTION 3: $\delta > \max\left[\underline{\delta}_0, \underline{\delta}_1\right]$

where $\underline{\delta}_0 := \frac{\Pi_3 - \kappa}{\Pi_3 - \varepsilon}$ and $\underline{\delta}_1 := \frac{-\beta_1 + \left(\beta_1^2 - 4\beta_0\beta_2\right)^{\frac{1}{2}}}{2\beta_0} < 1$.[19] It is essential to note that these conditions will always be satisfied when the time period converges to 0. Therefore, no assumption needs to be imposed in the continuous version of our game. We formalize all the previously discussed ideas in the following proposition.

[18] We show in the next paragraph that a necessary condition for this effect to be positive is that the time period be small enough.

[19] Where $\beta_0 = 2(\Pi_2 - \Pi_3)$, $\beta_1 = -3(\Pi_2 - \Pi_3) + \left(\underline{\kappa} - \varepsilon\right)$ and $\beta_2 = (\Pi_2 - \Pi_3)$. In the proof of Theorem 1, we show that $\underline{\delta}_1 < 1$.

PROPOSITION 3: *Under assumptions 1-3, for any* $\Delta \in \mathbb{R}_{++}$*, there exists a unique symmetric subgame perfect equilibrium* $\psi^* \in (0,1)$ *such that* $\mu(\psi^*) = 0$.

PROOF. *See Appendix.*

Proposition 3 expresses our main message in the case where the innovator faces two potential imitators. With some positive probability, imitators will strategically delay their entry on the market. The innovator might therefore retain monopoly profits for some time after discovery, even in the absence of intellectual property rights protection. This endogenous delay is driven by the dynamics of the market for technology. The imitators anticipate paying a smaller price for the innovation if they enter after their competitor as they know the innovator and the first entrant will compete aggressively to provide a license. Both imitators thus have an incentive to delay entry.

It is instructive to measure more specifically the consequences on expected profits and delay in the limit case where the length of the period converges to zero: $\Delta \to 0$, what we refer to as the continuous version of our game. We note that as $\Delta \to 0$, we have $\delta \to 1$, and therefore

$$\lim_{\Delta \downarrow 0} \upsilon_l := \upsilon_B = \Pi_3 - \kappa < \Pi_3 - \varepsilon = \lim_{\Delta \downarrow 0} \upsilon_f$$

At the limit we therefore obtain the classical form of a war of attrition with complete information (studied for instance by Hendricks, Weiss and Wilson, 1988 and Fudemberg and Tirole, 1991).[20] We further characterize the subgame perfect equilibrium in the limiting case and describe the expected profits of the innovator.

PROPOSITION 4: *Under Assumption 2, as* $\Delta \to 0$*, the limiting distribution of entry times is an exponential distribution with hazard rate*

$$\lambda = \frac{r\upsilon_l}{(\upsilon_f - \upsilon_l)} = \frac{r(\Pi_3 - \kappa)}{\kappa - \varepsilon}$$

Furthermore, the innovator's equilibrium expected profits are:

$$V_s = \tfrac{\pi_1}{r+2\lambda} + \tfrac{2\lambda}{r+2\lambda}\left[\Pi_3 + (\kappa - \varepsilon)\right]$$

PROOF. See Appendix.

The first part of Proposition 4 shows that the limit distribution of entry time t of each imitator, conditional on the other imitator not having entered before t is an exponential distribution $F(t) :=$

[20] At the limit, $\upsilon_f > \upsilon_l = \upsilon_B$ for all $t \in T$; which is the classical case examined in the theoretical analysis of war of attrition games. Note however that in the discrete version of the game, $\upsilon_l \neq \upsilon_B$. This difference motivated our choice to express the game in its discrete version and examine the limit as the time period converged to 0.

$1 - e^{-\lambda t}$. Furthermore, the second part of the proposition determines the expected profits of the innovator in such an environment. Until entry by one of the imitators, the innovator retains monopoly profits. After imitation by one of the firms, the competitor immediately follows and the innovator collects triopoly profits thereon. Finally, she also gathers licensing revenues from the first entrant.

We conclude this section by comparing the expected profits of the innovator when a market for technology exists to the profits she can expect if licensing is not feasible. Proposition 1 indicates that in the absence of a market for technology, entry occurs immediately and the expected profits of the innovator are Π_3. We can therefore express the net benefits for the innovator of an efficient market for technology as:

$$V_s - \Pi_3 = \underbrace{\frac{[\pi_1 - \pi_3]}{(r+2\lambda)}}_{\substack{Rewards \\ from\ delay}} + \underbrace{\frac{2\lambda}{(r+2\lambda)}[\kappa - \epsilon]}_{\substack{+\quad Licensing \\ revenue}}$$

The innovator benefits in two ways from a well functioning market for ideas: she collects licensing revenues but more importantly the dynamics on this market encourage the imitators to delay entry, thus preserving monopoly rents for a longer period of time.

3.4 Comparative Statics and Numerical Examples

In this section we discuss some comparative statics on the profits V_s of the innovator.

COROLLARY 2: *The profits of the innovator converge to monopoly profits when* $\kappa \to \Pi_3$

As the imitation cost converges to triopoly profits, the rents of the innovator converge to monopoly. Thus even in the absence of intellectual property rights protection, the innovator retains full "natural" protection. The intuition is that the rents of the first entrant converge to zero when the imitation cost converges to triopoly profits. The follower however retains strictly positive profits. Therefore the incentives to wait increase dramatically as $\kappa \to \Pi_3$ and the expected entry date converges to infinity thus allowing the innovator to enjoy monopoly rents for a very long period of time.

COROLLARY 3: *An increase in the imitation cost* κ increases *the equilibrium payoff of the innovator. An increase in the transfer fee* ϵ *decreases the equilibrium payoffs.*

PROOF. *See Appendix.*

Although seemingly intuitive, the reasoning underlying the previous results is not straightforward. An increase in the imitation cost κ has three distinct effects. First it increases the difference between κ and ϵ and thus further delays entry by imitators hoping for a reduced licensing fee. As κ increases, the innovator can thus retain monopoly profits for a longer period of time. Second, it increases the instantaneous value of licensing revenues $\kappa - \epsilon$. However, there is a third countervailing effect on the discounted value of licensing revenues. As imitators further delay entry, the licensing revenues are obtained at a later period. Overall, we show that the first effect will dominate the third, i.e the benefits from obtaining monopoly profits for an extra period of time dominates the cost of delaying licensing revenues. Note that if this was not the case, it would be optimal for the innovator to reduce slightly the licensing fee to obtain these revenues earlier. Overall, an increase in the imitation cost thus increases the revenues of the innovator. The intuition for the result on the transfer fee is identical except that the effects are reversed.

It is useful to illustrate these arguments in a particular example. We consider the case where demand is linear $p = a - q$, firms produce at a constant marginal cost c and compete a la Cournot on the product market. We fix the value of different parameters and in particular the transfer fee $\epsilon = 0.01$.[21] Given our assumptions, we can calculate the discounted sum of triopoly profits $\Pi_3 = 0.4$. We vary the imitation cost κ between ϵ and Π_3. We report the expected delay in entry and the ratio of profits (in our case without IP rights but a well functioning market for technology) over monopoly profits. In the last three columns we decompose the percentage contributions of the different revenue streams for the innovator in the absence of property rights. The first represents the percentage coming from monopoly profits before entry ($\frac{\pi_1}{r+2\lambda}$). The second represents the percentage coming from triopoly profits after entry ($\frac{2\lambda}{r+2\lambda}\Pi_3$) and finally the last represents the percentage obtained from licensing revenues ($\frac{2\lambda}{r+2\lambda}(\kappa - \varepsilon)$)

κ	Expected delay	Profits/Monopoly	% before entry	% after entry	% licensing
0.02	0.13	0.27	5	93	2
0.05	0.57	0.31	17	75	6
0.1	1.5	0.4	33	55	12
0.15	2.8	0.48	45	40	14
0.2	4.7	0.56	56	30	14
0.25	8	0.67	65	22	13
0.3	14	0.77	77	13	10
0.35	34	0.88	88	6	6
0.39	190	0.97	97	1	1
0.399	1945	1	100	0	0

[21] The market size is given by $a = 1$, marginal cost $c = 0.2$ and interest rate $r = 0.1$

In accordance with Corollary 3, as κ increases, the expected date of entry of the first imitator and the profits of the innovator increase. Furthermore, in accordance with Corollary 2, the ratio of profits without IP to monopoly profits converges to one as κ converges to Π_3. It is interesting to discuss the effects reported in last three columns. The percentage of overall profits coming from the monopoly position before entry naturally increases with κ. Indeed an increase in κ delays entry and therefore extends the period of time during which monopoly profits are collected. Conversely the percentage of overall profits coming from triopoly after entry decreases with the imitation cost. The more interesting result relates to licensing revenues. As κ increases, the instantaneous licensing revenue $\kappa - \epsilon$ increases. This effect is linear in κ. However, there is a countervailing effect: as κ increases, these revenues are obtained at a later date. Given that the effect of κ on delay is non linear it dominates as κ converges to Π_3. Therefore, the discounted value of expected licensing revenues initially increases with κ and then decreases as the effect of the delay starts dominating.

4 Free entry

The proponents of strong intellectual property protection generally argue that without such protection, imitators will freely enter the market and compete away the innovator's rents. The mechanism we unveiled in the previous section challenges this line of thought. We found the striking result that when contracting is possible, even in the absence of such protection, imitators will strategically delay entry, leaving substantive rents to the innovator. However, we restricted ourselves to the case of two potential imitators. In a concern for completeness, we need to examine whether our mechanism is still relevant when we consider free entry in the imitation market. We will find under certain conditions an even more striking result: firms, realizing that competition on the product market will be extremely fierce, might abstain from entering the market altogether thus leaving the innovator with full monopoly rents as if intellectual property rights existed.

The benchmark model is modified to account for free entry. At the beginning of each period the innovator and all the imitators that previously entered post a list of offered contracts. The free entry condition guarantees that all contracts that provide a positive expected value in equilibrium will be accepted. Note that if there was a limited number of potential entrant, they might turn down a contract providing positive profits if they expected in equilibrium to receive more attractive offers in the future. This is ruled out by the free entry condition. We also assume that regardless of the mode of entry (imitation or contracting) the entrants will compete on the market for technology.

It is useful to introduce some pieces of notation. We let K be the number of firms such that $\Pi_{K+1} \leq \kappa \leq \Pi_K$. Note that if contracting is not possible, $K-1$ imitators will immediately enter the

market and all players will earn profits Π_K. This summarizes the classical justification of IP protection. Similarly, we define L as the number of firms such that $\Pi_{L+1} \leq \epsilon \leq \Pi_L$

Our first result demonstrates that there will always exist an equilibrium where the innovator retains full monopoly profits as if IP protection existed.

PROPOSITION 5: There exists an equilibrium such that:

(i) If one imitator enters at t, $L-2$ imitators enter at $t+1$ by signing a contract with either the innovator or the imitator.

(ii) No imitator initially enters and the innovator obtains profits Π_1

PROOF. See Appendix.

The first part of Proposition 5 indicates that there always exists an equilibrium where $L-2$ imitators follow after entry of an initial imitator. In this equilibrium the innovator and the lead imitator both offer $L-2$ licenses at a price of ϵ. Given that the competitor follows this strategy, there is no profitable deviation as entry will occur regardless of the strategy adopted. The second part of the proposition then naturally follows. Given that competition on the market for technology is so intense, entrants know that future profits Π_L will not be sufficient to cover the initial imitation cost κ.

We learn from Proposition 5 that there exists an equilibrium such that the innovator retains full monopoly rents. However, this is one among several possible equilibria. Other equilibria can guarantee higher profits for the innovator and the lead imitator by constructing punishment strategies. Note that the most sever punishment is to revert to the equilibrium described in Proposition 5 where the profits are reduced to Π_L. It is therefore useful to determine sufficient conditions that guarantee that the behavior described in Proposition 5 is indeed the unique equilibrium outcome. We will show that this will be the case under the following assumption on profits:[22]

Assumption A: $\quad \forall R \leq K \quad \Pi_R < (L-R+1)\Pi_L$

The interpretation of this assumption will become clear when we explain the intuition of Proposition 6. It is useful to provide a sense of conditions under which this assumption will be satisfied. Consider, as we did in the previous section, a linear demand $p = a-q$ for the good, with a constant marginal cost of production c and Cournot competition. Note that the condition will be independent of the values of a and c. In this case the assumption can be expressed in the following way: $\Pi_K < (L-K+1)\Pi_L$. The following table indicates for different values of K, the range of values of L that guarantee that the assumption is satisfied.

[22] Note that more restrictive assumptions might be sufficient to guarantee unicity

K	Range of L
2	3-5
3	4-11
5	6-29
10	11-109

This table can be interpreted in the following way: if without licensing, ten firms enter the market ($K = 10$) then the range of values for L such that the assumption is satisfied is large ($L \in [11 - 109]$). In other words unless the transfer cost is extremely small, the assumption will be satisfied. To put it in perspective, in this case if the transfer cost ϵ is greater than one percent of the imitation cost κ, this will be the case. We note that in general the condition is more easily met when K is large (i.e when the imitation cost is small). For the case of linear demand and Cournot competition, the assumption will therefore generally be satisfied. Based on this assumption we can derive the following result.

PROPOSITION 6: Under assumption A, as $\Delta \to 0$*, the unique equilibrium outcome is that the innovator retains monopoly profits* Π_1

PROOF. See Appendix.

Proposition 6 states that under assumption A, the unique equilibrium outcome is that the innovator retains monopoly rents as if intellectual property rights were protected. A natural protection emerges. The role of Assumption A is to guarantee that in equilibrium, the number of licenses signed is exactly L. To understand this result, consider an equilibrium where $R \in [L, K]$ licenses are signed. Assumption A guarantees that there is a profitable deviation in this case. After the R licenses have been signed, consider the following deviation by the innovator: offer $(L-R)$ licenses at a price of Π_L. Given the free entry assumption these contracts will be accepted by potential entrants. Furthermore, in any subgame perfect equilibrium, strategies need to form a Nash Equilibrium in all subgames. The worst payoff the innovator can obtain in any subgame is Π_L. Therefore if $\Pi_R < (L - R)\Pi_L + (1 - \delta)\Pi_L + \delta\Pi_L$ the deviation is profitable (the first term are the licensing revenues, the second the period payoff and the third the worst punishment that can be imposed). Assumption A guarantees that this is the case for all $R \in [L, K]$ and that the deviation is therefore profitable.

Assumption A that guarantees the unicity of the equilibrium where the innovator retains monopoly rents constrains the competition on the product market to be weak. Paradoxically this creates intense competition on the market for technology. The relation between these two markets is intricate and generates important dynamics that we highlight in this paper.

5 Robustness checks

5.1 Royalty rates

We restricted the analysis in the previous sections to contracts involving only fixed fees. We now allow for more elaborate contracts, involving both fixed fees and royalty rates. We return to the case of two potential imitators as in the benchmark model. To determine the optimal contracts with royalty rates, we need to impose more structure on demand and competition. We suppose the innovation is a process innovation which allows production of the good at a constant marginal cost c. We suppose that the imitators if they do not copy or sign a contract, earn zero profits on the product market. The competitors face a linear demand $p = a - q$ for the good and the innovator competes a la Cournot with the imitators who acquired the invention.

To find the Subgame Perfect Nash Equilibrium of this game we start by determining the outcome of the game starting after one imitator has obtained the invention. We suppose that the first imitator, that we call the leader, has acquired the invention through the contract (F, ρ_l).

LEMMA 4: For all values of ρ_l the follower will enter immediately and sign a contract with no royalty rate and a fixed fee equal to ϵ

PROOF. See Appendix.

We find the surprising result that royalty rates will not be used in equilibrium. It is well known that the royalty rate can increase the joint surplus in a duopoly as it serves as a commitment to restrict production. However, in our triopoly case, some of the rents go to the third player. If for instance a bilateral contract including a royalty rate is signed between the innovator and the follower, the royalty rate will force the licensee to restrict supply yielding rents to the innovator. However, the strategic response of the lead imitator is to increase production given that he expects the price to be higher.[23] This will further decrease the profits of the follower. In equilibrium the royalty rate is chosen to maximize the joint bilateral surplus between the innovator and the follower.[24] The rents obtained by the innovator are not sufficient to cover the losses of the licensee as rents are dissipated by the strategic response of the competitor. The optimal choice is therefore to impose no royalties.

The same arguments will also imply that the innovator will not include a royalty rate in the contract she proposes to the lead innovator as she knows that some of the rents will be appropriated by the

[23] We show in the proof that the optimal quantity produced by the lead imitator is increasing in the royalty rate imposed on the follower.

[24] If that was not the case, the licensor could deviate and propose a different contract with the optimal royalty rate and a higher fixed fee leaving the licensee indifferent. The licensor would thus increase its profits by appropriating some of the higher joint surplus. Note that this does not depend on the strategic response of the competitor.

follower. The game turns out to be equivalent to the one we studied in section 3. This is the conclusion of the following proposition.

PROPOSITION 7: When contracts can include fixed fees and a royalty rate, there is a unique symmetric SPE in which all the contracts impose zero royalty rates and the equilibrium entry strategy of the imitators, G, is the exponential distribution of parameter λ given in section 3:

$$G(t) = 1 - e^{-\lambda t} \text{ for } t \in [0, \infty)$$

PROOF. See Appendix.

The results of Section 3 on delay and innovator's rents are robust to the introduction of royalty rates. It also seems that this logic also applies to the case of free entry. We believe that this contributes to the literature on the shape of contracts that has examined in a number of situations the choice between fixed fee and royalty rates. We elaborate on the links with the existing literature in the next section.

6 Links with the literature

One of the contributions of our paper is to show how endogenous delay in imitation can emerge due to the dynamics in the market for technology. Other explanations have been proposed in the literature for such delays. For instance Scherer suggests that technological constraints can generate "natural lags" in imitation. This explanation however does not depend on strategic responses of firms involved but is a direct consequence of the characteristics of the environment. Benoit (1985) considers an environment where the value of a non-patentable innovation is unknown. The imitator might prefer to wait before entering to obtain better information on this value. The author shows that an increase in the cost of innovating[25] can then paradoxically increase the chances that an innovation will be undertaken. Indeed an increase in the cost increases the incentives to wait and thus generates higher rents for the innovator. These extra profits can under certain conditions dominate the increase in costs. Choi (1998) considers imitation in a world with imperfect patents. The patent holder can decide to initiate an infringement suit if he observes imitation. This is however a risky procedure that can lead to the patent being declared invalid and thus allow further entry. This gives rise to a war of attrition where imitators have incentives to let competitors enter before them and face the risk of a lawsuit that will reveal the validity of the patent.

We propose an alternative explanation for endogenous delays in imitation purely based on the dynamics of the market for technology. Arora et al. (2002) emphasize the growing importance of the

[25] The cost of producing the innovation is supposed to be the same for the innovator and the imitator

markets for technology and we build on this evidence. Note that this explanation does not rely on the existence of patents as in Choi. Furthermore, in our paper, we show that an increase in the number of potential imitators can have a positive effect on the incentives to innovate. In the case of free entry, we show in particular that the initial innovator might retain full monopoly rents.

Our mechanism rests on the dynamics of the market for technology. An important paper by Arora and Fosfuri (2003) point out that it can be optimal for a producer to license her technology to a potential imitator. They argue that under certain circumstances, the revenues the firm can expect from licensing to another competitor dominates the loss due to the erosion of overall industry profits. They determine the equilibrium number of licenses in a model where two competitors simultaneously determine their licensing behavior. The trade-off they consider is similar to the one that gives rise to Assumption 2 in section 3.[26] If the profits do not decrease two fast with the number of players, competition on the market for technology will tend to be intense. Paradoxically in our context, we show that this fierce competition has a positive effect on the rents the innovator can expect.

Another branch of the literature has studied the problems that can arise when an inventor and a producer sign a licensing contract in the absence of property rights. Anton and Yao (1994) study an environment where a financially constrained inventor, who cannot bring his innovation to the market, can potentially license to two producers. Two problems arise when the parties attempt to sign a contract. First, only the inventor is informed about the value of the innovation, thus making the producer wary of signing a contract. Second, if the inventor does reveal his information the producer can then fully appropriate the invention without any form of payments. The authors show that under certain conditions the parties can overcome this problem and the inventor can still obtain some rents. The mechanism guaranteeing that outcome starts with full revelation by the inventor. The producer will still subsequently sign a contract under the threat that the inventor will also reveal the idea to the competitor if he decides to copy without appropriate payments.[27] In another paper, Anton and Yao (2002) propose a different mechanism based on partial disclosure of the idea and the issuance of a bond that allows the inventor to appropriate some of the returns from his invention.[28] In our paper we do not consider these informational issues. Including them could be an interesting extension of our work. However, we believe that these effects are less relevant in the problems we consider. Indeed, the inventor in our setup has already produced a final version of the invention and is using it on the product market, as opposed to the inventor in Anton and Yao who cannot translate her invention into

[26] Consider Assumption 2 for $\epsilon = 0$ $2\Pi_3 \geq \Pi_2$. In a static framework this condition guarantees that starting from a duopoly, if the competitor does not provide a license, it is beneficial for a firm to provide a license to a potential entrant for a price of Π_3.

[27] Conditions guarantee that this threat is suffiicient to remove the temptation of copying without payments.

[28] The amount of self exposure to expropriation through disclosure and through the bond, signals the value of the invention.

a product. It therefore seems reasonable to assume that all imitators observe fully the value of the invention from the success on the market and that adverse selection issues are not essential.

Finally, our robustness checks on the shape of licensing contracts contributes to the literature on optimal licensing. There is an extremely vast literature on the subject. Kamien and Tauman (1986) show for instance that fixed fee licensing dominates pure royalty licensing. Shapiro (1985) demonstrates that a licensing contract that combines fixed fee and royalty rates allows the licensor to reproduce monopoly profits by changing the marginal cost of production of the competitor. Sen and Tauman (2006) consider general contracts with a combination of fixed fee and royalty rate and compare the case where the innovator is an insider to the case where she is an outsider.[29] They show that positive royalty rates will always be used for sufficiently valuable innovations. To the best of our knowledge, in most of the papers in this literature, the innovator moves first and offers a collection of contracts. In our framework, the contracts are negotiated bilaterally and sequentially. We show that in the particular case we consider, positive royalty rates will never be used. The intuition for this result reported in section 5 seems quite general. This is an interesting finding that to the best of our knowledge is new in the literature and merits further investigation to determine the general conditions under which it holds.

7 Conclusion

In this paper, we examine the profits of an innovator and the timing of entry of potential imitators in a world without intellectual property rights protection. If a market for technologies does not exist, imitators enter immediately and compete away the innovator's rents. However, when we introduce a well functioning licensing market, we show that endogenous delay in entry occurs. The imitators know that when one of them enters, he will compete with the innovator to provide a license thus reducing the cost of entry. This will naturally generate a war of attrition where follower imitators earn higher profits then leaders and where therefore everyone wishes to wait. Furthermore, when we consider free entry, we show that under certain conditions, the unique equilibrium is one where the innovator retains full monopoly rents. Indeed competition on the market for technology can be so fierce that the final profits an imitator can expect do not cover the initial imitation cost. From the point of view of the initial imitator, the dynamics of the market for technology leads to excessive entry.

The condition that guarantees unicity of the equilibrium constrains profits on the product market. It is necessary that the rate at which profits decrease with the number of competitors on the market be slow. The condition therefore requires that the product market not be very competitive. Paradoxically this implies that the licensing market will be very competitive. Intuitively, if the rents are dissipated

[29] An insider is defined as an inventor who is also a competitor on the product market

at a low rate when a new entrant is accepted, licensing becomes a more attractive option. This demonstrates that taking into account the interactions between the product market and the market for technology is essential.

The results of this paper challenge the traditional view on the necessity of intellectual property protection. As we pointed out in the introduction, we considered a framework where information asymmetries are small as the value of the innovation is publicly observed through its performance on the market. However, intellectual property rights could still be important to facilitate the licensing of more basic innovations. This could be an interesting extension of our model. How would a basic idea be licensed to a future producer in the presence of information asymmetries and given that the producer will face an efficient market for technologies?

Finally we want to point out that we left aside the question of social welfare. This is the natural next step in this research agenda attempting to take into account the dynamics of the market for technology. We have shown that the possibility of contracting will generate a natural degree of protection for the innovator. A number of question arise. Will this protection be excessive or sufficient? Will there be a systematic link between the degree of natural protection and the underlying value of the innovation? Should entry in the licensing market be regulated by the decision maker to adjust the incentives to innovate? All these fascinating questions will be the object we hope of future work.

8 Bibliography

ANTON, J. and D. YAO (1994), "Expropriation and Inventions: Appropriable Rents in the Absence of Property Rights", *The American Economic Review*, 84, 190-209.

ANTON, J. and D. YAO (2002), "The Sale of Ideas: Strategic Disclosure, Property Rights and Contracting", *The Review of Economic Studies*, 69, 513-531.

ARORA, A., FOSFURI, A. and A. GAMBARDELLA (2002), "Markets for Technology: the Economics of Innovation and Corporate Strategy", *The MIT Press*

ARORA, A., FOSFURI, A. (2003), "Licensing the Market for Technologies", *Journal of Economic Behavior and Organization*, 52, 277-295.

BENOIT, J-P (1985), "Innovation and Imitation in a Duopoly", *Review of Economic Studies*, 52, 99-106.

BERNHEIM. D. (1984), "Strategic Deterrence of Sequential Entry into an Industry", *The Rand Journal of Economics*, 15, 1-11.

CHOI, J. P. (1998), "Patent Litigation as an Information-Transmission Mechanism", *The American Economic Review*, 88, 1249-1263.

FUDENBERG, D. and J. TIROLE (1991), "Game Theory", *The MIT Press*.

HEINDRICKS, K., WEISS, A. and C. WILSON (1988), "The War of Attrition in Continuous Time with Complete Information", *International Economic Review*, 29, 663-680.

SCHERER, F. M. (1980), "Industrial Market Structure and Economic Performance"

SEN, D. and Y. TAUMAN (2006), "General Licensing Schemes for a Cost-reducing Innovation", *Games and Economic Behavior*, 59, 163-186.

SHAPIRO, C. (1985), "Patent Licensing and R&D Rivalry", *The American Economic Review*, 75, 25-30.

9 Appendix

PROOF OF PROPOSITION 1

In the case without licensing, the game is dominance solvable. Imitating at $t = 0$ is a dominant strategy for both players. Any entry time $t' > 0$ leads to strictly lower profits, regardless of the strategy of the competitor as it abandons the profits in the initial period without affecting the profits after entry or the cost of entry. The equilibrium specified in Lemma 1 is therefore the unique SPNE. Therefore, immediately after innovation, the flow profits of the innovator are reduced to π_3 and his expected profits are $\frac{\pi_3}{r} = \Pi_3$ and the profits of the imitators are $\Pi_3 - \kappa$. □

PROOF OF LEMMA 1

We first consider the case where imitator i entered by contracting. Consider an equilibrium of the subsequent subgame starting at $t + 1$. Let $\tau \geq t + 1$ be the date at which a contract is signed in equilibrium. Note that this date τ has an upper bound given that imitator j has an alternative strategy which is to enter immediately at $t + 1$ by copying.

Step 1: *The price of a license at τ is ϵ*

Suppose that this is not the case, and the license is sold at a price $p_\tau > \epsilon$, for instance between firm i and an entrant. Then an optimal deviation for the innovator is to offer a slightly lower price. She will thus raise licensing revenues without changing the continuation value of the game.

Step 2: *Under assumption 2 all license will be signed at $t + 1$ (immediately after entry of the first imitator)*

Suppose $\tau > t+1$. We want to show that there is it is optimal to deviate and offer an extra license at time $\tau - 1$. Given the result in step 2, at τ the license will be signed for a price $p_\tau = \epsilon$. If the innovator deviates and offers a license at time $\tau - 1$ we determine the maximum price that can be charged. Firm j by signing at $\tau - 1$ for a price $p_{\tau-1}$ expects profits $(1 - \delta)\Pi_3 + \delta\Pi_3 - p_{\tau-1}$. If it rejects the license, it will obtain profits $\delta\Pi_3 - \delta\epsilon$. The maximum price that can be charged is therefore $p_{\max} = (1 - \delta)\Pi_3 + \delta\epsilon$. The deviation that consists in offering the license at $\tau - 1$ at a price of $p_{\max}$ is therefore profitable if $(1 - \delta)\Pi_3 + \delta\Pi_3 + p_{\max} - \epsilon \geq (1 - \delta)\Pi_2 + \delta\Pi_3$. This condition is equivalent to Assumption 2: $2\Pi_3 - \epsilon \geq \Pi_2$. If assumption 2 is satisfied this cannot be an equilibrium.

Step 3: *From step 1 and 2 we can immediately conclude that if imitator i entered by contracting, imitator j will immediately follow and pay a minimal price for the license.*

Finally, we show that if imitator i entered by imitating and not by contracting, imitator j will immediately follow. At period $t + 1$ imitator j has to decide whether to enter or not. If he enters his expected profits are $\Pi_3 - p_{t+1}$ where p_{t+1} is the price on the market for a license. If he delays entry, his

expected profits are $\delta(\Pi_3 - p_{t+2})$. However the game in period $t+2$ if he delays entry is identical to the game in $t+1$. Therefore, in all subgame perfect equilibria, given that the innovator is the only other active player, we will have $p_{t+1} = p_{t+2}$. Therefore imitator j enters in period $t+1$ to avoid forgoing the triopoly profits for one extra period. □

PROOF OF LEMMA 2

Setting a price $p_t > \kappa$ is a dominated strategy. If an imitator decides to enter, he will do so by copying as according to the results of Lemma 1, his licensing revenues will be zero. Therefore offering a license at a price $p_t = \kappa$ would yield higher revenues as some licensing revenues would be obtained. If no imitator enters, the payoffs are identical in the two cases.

We also show that setting a price $p_t < \kappa$ is a dominated strategy.

- either an imitator will enter at t if $p_t = \kappa$: then by offering $p_t < \kappa$ the innovator looses licensing revenues

- either an imitator will not enter at t if $p_t = \kappa$: then by offering $p_t < \kappa$ the imitator might enter. Given that $p_t < \kappa < \Pi_3$ this would decrease overall profits.

By deletion of dominated strategies, the only equilibrium is to set $p_t = \kappa$ at every period. □

PROOF OF PROPOSITION 2. the results of Proposition 2 are a direct consequence of Lemma 1 and Lemma 2.

PROOF OF PROPOSITION 3.We follow four steps. First, we show that equation (2) $\mu(\psi) = 0$ has a solution $\psi^* \in (0,1)$. Second, we prove that ψ^* is unique. Third, we argue that the stationary symmetric strategy pair ψ^* constitute a Nash equilibrium. And finally, we show that the Nash equilibrium is, indeed, perfect.

Step 1. From equation (2) we know that $\mu(\psi) = \psi[\upsilon_f - \upsilon_B] + (1-\psi)[\delta V(\psi) - \upsilon_l]$. $\mu(\psi)$ is a continuous function of ψ. We have $\mu(0) = \delta V(0) - \upsilon_l = -(1-\delta)\upsilon_l < 0$ and $\mu(1) = (\upsilon_f - \upsilon_B) = \delta(\Pi_3 - \varepsilon) - (\Pi_3 - \kappa) > 0$, because by Assumption 3, $\delta > \frac{(\Pi_3 - \kappa)}{(\Pi_3 - \varepsilon)}$. Thus by continuity of $\mu(\psi)$ there exists at least a $\psi^* \in (0,1)$ such that $\mu(\psi^*) = 0$.

Step 2. We show uniqueness by demonstrating that $\forall \psi \in (0,1) : \mu(\psi)$ is a strictly increasing function so that the solution of $\mu(\psi) = 0$ is unique.

We show that

$$\frac{d\mu(\psi)}{d\psi} = (\upsilon_f - \upsilon_B) - (\delta V(\psi) - \upsilon_l) + (1-\psi)\delta\frac{dV(\psi)}{d\psi}$$

Computing $\frac{dV(\psi)}{d\psi}$ and using equation (1) it follows that

$$\frac{d\mu(\psi)}{d\psi} = (\upsilon_f - \upsilon_B + \upsilon_l(1-\delta) - \delta(\upsilon_l - \upsilon_B)) + 2\delta(\upsilon_l - \upsilon_B)\psi$$

Because $(\upsilon_l - \upsilon_B) > 0$, a sufficient condition for $\frac{d\mu(\psi)}{d\psi} > 0 \ \forall \psi > 0$ is that $\frac{d\mu}{d\psi}(0) > 0$.

We have:

$$\frac{d\mu}{d\psi}(0) := \mathcal{Z}(\delta) = 2(\Pi_2 - \Pi_3)\delta^2 - (3(\Pi_2 - \Pi_3) - (\kappa - \varepsilon))\delta + (\Pi_2 - \Pi_3)$$

Observe that: (i) $\mathcal{Z}(0) = (\Pi_2 - \Pi_3) > 0$; (ii) $\mathcal{Z}(1) = (\kappa - \varepsilon) > 0 : \forall \kappa \in \left[\underline{\kappa}, \Pi_3\right]$; and (iii) $\mathcal{Z}(\delta)$ is a strictly convex function of δ $(\mathcal{Z}''(\delta) = 4(\Pi_2 - \Pi_3) > 0)$

If $\min_{\delta \in (0,1)} \mathcal{Z}(\delta) > 0$ our proof is concluded

If $\min_{\delta \in (0,1)} \mathcal{Z}(\delta) < 0$ we need to impose conditions on δ that guarantee that $\frac{d\mu}{d\psi}(0) > 0$. Consider the largest solution to the quadratic equation $\mathcal{Z}(\delta) = 0$. We know that this largest root is given by:

$$\underline{\delta}_1 = \frac{-\beta_1 + \sqrt{\beta_1^2 - 4\beta_0\beta_2}}{2\beta_0}$$

where

$$\beta_0 = 2(\Pi_2 - \Pi_3); \beta_1 = -(3(\Pi_2 - \Pi_3) - (\kappa - \varepsilon)); \beta_2 = (\Pi_2 - \Pi_3)$$

Finally, because by Assumption 3, $\delta > \underline{\delta}_1$and $\mathcal{Z}(\delta)$ is strictly convex and $\mathcal{Z}(1) > 0$, we can conclude that $\mathcal{Z}(\delta) > 0$.

Step 3. Notice that the candidate strategy for a stationary symmetric equilibrium is: *always* $\psi^* \in (0, 1)$. To show that the pair *always* ψ^* is a Nash equilibrium, it suffices to prove that each imitator is indifferent between *all* time periods $t \in T := \{0, 1, 2...\}$ to implement the innovation. Observe that we can write $\mu(\psi)$ as

$$\mu(\psi) = \alpha_0\psi^2 + \alpha_1\psi + \alpha_2$$

where $\alpha_0 := [\delta(\upsilon_l - \upsilon_B)]$; $\alpha_1 := [\upsilon_f - \delta\upsilon_l + (\upsilon_l - \upsilon_B)(1 - \delta)]$ and $\alpha_2 := -\upsilon_l(1 - \delta)$. Hence, ψ^* is

$$\psi^* = \frac{-\alpha_1 + \sqrt{\alpha_1^2 - 4\alpha_0\alpha_2}}{2\alpha_0}$$

Now suppose that imitator $a \in \mathcal{I}$ chooses *always* ψ^*, then the value of implementation for imitator $b \in \mathcal{I}$ at each possible implementation time is

$$V_b(\psi^*) := \psi^*\upsilon_B + (1 - \psi^*)\upsilon_l$$

the same $\forall t \in T := \{0, 1, 2...\}$ given the stationary of the payoffs.

Step 4. To demonstrate that *always* ψ^* is indeed a perfect equilibrium it is sufficient to notice that, because of the stationary of the payoffs, all subgames in which the imitators are still deciding

when to implement the innovation have the same structure (i.e., they are isomorphic) as the original implementation game. Hence by Step 3, *always* ψ^* is indeed perfect. □

PROOF OF PROPOSITION 4.

We follow two steps. First, we show that when $\Delta \to 0$, the limit of the equilibrium solution also goes to zero, that is $\lim_{\Delta \downarrow 0} \psi^*(\Delta) = 0$. In the second step we determine the limit of the ratio $\frac{\psi^*(\Delta)}{\Delta}$.

Step 1. From Theorem 1, we know that $\mu(\psi) = \alpha_0\psi^2 + \alpha_1\psi + \alpha_2$ and because $\alpha_j = \alpha_j(\Delta)$ for $j = 0, 1, 2$ it follows that $\psi^*(\Delta)$. Because when when $\Delta \to 0$: (i) $\delta \to 1$; and (ii) $\upsilon_l \to \upsilon_B$; it follows directly that: (i) $\alpha_0 \to 0$; (ii) $\alpha_1 \to \kappa - \varepsilon$; and (iii) $\alpha_2 \to 0$. Therefore because in equilibrium it must be that

$$\lim_{\Delta \downarrow 0} \mu(\psi^*) = \lim_{\Delta \downarrow 0} \left(\alpha_0 \psi^{*2} + \alpha_1 \psi^* + \alpha_2\right) = (\upsilon_f - \upsilon_l) \lim_{\Delta \downarrow 0} \psi^*(\Delta) = 0$$

which implies that $\lim_{\Delta \downarrow 0} \psi^*(\Delta) = 0$.

Step 2. We want to determine the following limit: $\lim_{\Delta \downarrow 0} \frac{\psi^*(\Delta)}{\Delta}$. We therefore consider a Taylor expansion of the function ψ^*.

More specifically, consider the following function: $f(x) = -\alpha_1 + \sqrt{\alpha_1^2 + x}$. Given that $\lim_{\Delta \downarrow 0} \alpha_0\alpha_2 = 0$ we can consider a Taylor expansion of $f(x)$ at 0.

$$f(-4\alpha_0\alpha_2) = f(0) + f'(0)(-4\alpha_0\alpha_2) + o(\Delta^2) = \frac{1}{2\sqrt{\alpha_1^2}}(-4\alpha_0\alpha_2) + o(\Delta^2)$$

Therefore: $\frac{\psi^*(\Delta)}{\Delta} = \frac{-\alpha_2}{\alpha_1 \Delta} + o(\Delta) = \frac{1-\delta}{\Delta}\frac{\upsilon_1}{\alpha_1} + o(\Delta)$

To conclude the proof, we have $\lim_{\Delta \downarrow 0} \frac{(1-\delta)}{\Delta} = r$ and $\lim_{\Delta \downarrow 0} \alpha_1 = \kappa - \varepsilon$, so we finally obtain $\lim_{\Delta \downarrow 0} \frac{\psi^*(\Delta)}{\Delta} = \lambda = \frac{r(\Pi_3 - \kappa)}{\kappa - \varepsilon}$ □

PROOF OF CORROLARY 3

Totally differentiating $\upsilon_s(\theta)$ with respect to κ and considering that

$$\frac{d\lambda}{d\kappa} = \frac{-r(\kappa - \varepsilon) - r(\Pi_3 - \kappa)}{(\kappa - \varepsilon)^2} = -\frac{1}{\kappa - \varepsilon}\left[r + \frac{r(\Pi_3 - \kappa)}{\kappa - \varepsilon}\right] = -\frac{1}{\kappa - \varepsilon}[r + \lambda]$$

we have that

$$\frac{d\upsilon_s(\theta)}{d\kappa} = \frac{2(r + \lambda)}{(r + 2\lambda)^2}\left[\frac{\Pi_1 - \Pi_3}{\kappa - \varepsilon} - r\right] + \frac{2\lambda}{r + 2\lambda}$$

because the highest value that κ may assume is Π_3 and the lowest value that ε can take is zero, it follows that a sufficient condition for $\frac{d\upsilon_s(\theta)}{d\kappa} > 0$ for all imitation technologies is that

$$\Pi_1 - 2\Pi_3 > 0$$

This condition will be satisfied in all the markets that we consider. □

PROOF OF PROPOSITION 5

(i) Consider the following strategies for the innovator and the first imitator: at each date $\tau \geq t$ if $N < L$ firms hold the innovation, offer $L - N$ licenses at a price of ϵ.

The innovator and the imitator have no profitable deviations. Consider the case of the innovator (after entry they are in a symmetric position, so the reasoning is the same for the imitator):

- If she raises the price she charges for the licenses or reduces the number of licenses she offers, the potential entrants still enter but sign a license with her competitor.

- If she offers an extra license, it will not change the number of entrants as only L firms will enter

(ii) Given the first part of the proof, no imitator will want to incur the cost of imitation κ which is greater than profits Π_L. The optimal strategy for the innovator is to offer no licensing contract. Overall she will therefore maintain monopoly profits

PROOF OF PROPOSITION 6

The proof proceeds in a number of steps. In the first part of the proof we consider the subgame following entry by one of the imitators. We study competition on the market for technology. We first show that the total number of contracts signed in equilibrium can only be L. We then show that the price of licenses in equilibrium at the stage where the last contracts are signed has to be equal to ϵ. In step 3 we use these results to show that all contracts will be signed immediately after entry and we can then conclude that the only equilibrium leads to immediate entry of $L-2$ imitators. In part 2 we are able to conclude that the unique equilibrium is characterized by no entry and full monopoly rents for the innovator.

Part I: we start by considering the subgame following entry by one imitator at t.

Step 1: the number of contracts signed in equilibrium is L

Consider any equilibrium. Denote R the number of contracts signed in equilibrium. Let τ be the last date at which a license is signed in that equilibrium.

We can first point out that $R \leq K$: if $R > K$ then it would be profitable for firms to enter by copying and they would do so under the free entry condition. A profitable deviation would therefore be to offer more licenses and obtain licensing revenue.

Secondly $R \geq L$: in any subgame perfect equilibrium no contract would be offered at a price strictly lower than the transfer cost ϵ. Furthermore at a price higher than ϵ no firm would find it profitable to enter.

We now consider the cases where $R \in [L, K]$. Suppose that at $t' < \tau$ the innovator deviates and offers $(L - R)$ licenses at a price of Π_L. Given the free entry assumption these contracts will be accepted by potential entrants. Furthermore, in any subgame perfect equilibrium, strategies need to form a Nash Equilibrium in all subgames. The worst payoff the innovator can obtain in any subgame is Π_L. Therefore if $\Pi_R < (L - R)\Pi_L + (1 - \delta)\Pi_L + \delta\Pi_L$ the deviation is profitable (the first term are the licensing revenues, the second the period payoff and the third the worst punishment that can be imposed). Assumption A guarantees that this is the case for all $R \in [L, K]$.

Step 2: All licenses signed at τ need to be signed at a price of ϵ

Suppose that this is not the case, and a license is sold at a price $p > \epsilon$, for instance between firm i and an entrant. Then an optimal deviation for the innovator is to offer a slightly lower price. She will thus raise licensing revenues. Furthermore, from step 1 we know that at period τ , absent any deviation, L firms will have entered. This is already the minimal payoff in any subgame. Therefore the deviation cannot decrease the continuation payoffs. This deviation is therefore optimal and it cannot be the case that a license is sold in equilibrium at a price greater than ϵ in period τ.

Step 3: All license will be signed at t (immediately after entry of the first imitator)

Suppose $\tau > t$. We want to show that there is it is optimal to deviate and offer an extra license at time $\tau - 1$. We denote U the number of firms that are competing on the market at time $\tau - 1$. Given the result in step 2, at τ all licenses will be signed for a price $p_\tau = \epsilon$. If the licensee deviates and offers a license at time $\tau - 1$ we determine the maximum price that can be charged. The expected utility of the entrant is at least $(1-\delta)\Pi_{U+1} + \delta\Pi_L - p_{\max}$ given that Π_L is the worst case scenario for future payoffs. Given that $\Pi_L \geq \epsilon$, we therefore have $p_{\max} \geq (1-\delta)\Pi_{U+1} + \delta\epsilon$. The deviation that consists in offering one license at $\tau - 1$ at a price of $p_{\max}$ is therefore profitable if $(1-\delta)\Pi_{U+1} + p_{\max} - \epsilon > (1-\delta)\Pi_U$. Given that $p_{\max} \geq (1 - \delta)\Pi_{R+1} + \delta\epsilon$ a sufficient condition is therefore $2(1 - \delta)\Pi_{U+1} - (1 - \delta)\epsilon > (1 - \delta)\Pi_U$. Given Assumption 2, this condition will be satisfied and therefore signing a license at $\tau > t$ cannot be an equilibrium outcome.

Part II: We can conclude from Part I that the unique equilibrium outcome is that immediately after entry, $L - 2$ contracts will be signed at a price of ϵ. Therefore the expected profits of the first entrant are equal to Π_L. As in the proof of Proposition 4 we can conclude that no imitator will enter the market and that the innovator will retain monopoly rents.

LEMMA 4:

Consider the case where one imitator, that we call the leader, has acquired the invention through the contract (F, ρ_l).

Suppose the second imitator, called the follower, accepts the contract (G, ρ_f), we can then determine the quantities produced in the market. These quantities are solution to the following problems:

The innovator chooses $q_s = \arg\max_q\ [a - q - q_l - q_f]q - cq$

The lead imitator chooses $q_l = \arg\max_q\ [a - q - q_s - q_f]q - (c + \rho_l)q$

The follower chooses $q_f = \arg\max_q\ [a - q - q_s - q_l]q - (c + \rho_f)q$

The equilibrium quantities are: $q_s = \frac{a-c}{4} + \frac{\rho_l}{4} + \frac{\rho_f}{4}$, $q_l = \frac{a-c}{4} - \frac{3\rho_l}{4} + \frac{\rho_f}{4}$ and $q_f = \frac{a-c}{4} - \frac{3\rho_f}{4} + \frac{\rho_l}{4}$.

The contract proposed in equilibrium will always involve a royalty rate that maximizes the joint surplus of the licensor and licensee. Denote ρ^* the royalty rate that maximizes the surplus. Suppose the contract involves a different royalty rate ρ and a fixed fee F. The utility of the licensor in this case is $\pi_{licensor}(\rho)+F$ and the licensee $\pi_{licensee}(\rho)-F$. Suppose the licensor proposes a contract (ρ^*, F') with $F' = F + \pi_{licensee}(\rho^*) - \pi_{licensee}(\rho)$. This is accepted by the licensee who is indifferent between the two contracts. The utility of the licensor is then strictly higher: $\pi_{licensor}(\rho^*)+F' = \pi_{licensor}(\rho^*) + F + \pi_{licensee}(\rho^*) - \pi_{licensee}(\rho) > \pi_{licensor}(\rho)+F$ by definition of ρ^*. Thus necessarily the contract proposed in equilibrium involves a royalty rate that maximizes the joint surplus of the licensor and licensee. The fixed fee will then be used for transfers and will be determined by the competition between the innovator and the lead imitator to provide the license.

Consider the case where the equilibrium contract is signed by the innovator. The optimal royalty rate imposed by the innovator is solution to:

$\rho^* = \arg\max_{\rho_f}(\Pi_i + \Pi_f) = \arg\max_{\rho_f}(p - c)(q_i + q_f)$

Let $S = \Pi_i + \Pi_f = (a - Q - c)(q_i + q_f)$

We have $\frac{\partial(q_i+q_f)}{\partial\rho_f} = -\frac{1}{2}$ and $\frac{\partial(Q)}{\partial\rho_f} = -\frac{1}{4}$. Overall we have that $\frac{\partial S}{\partial\rho_f} = -\frac{1}{4}\rho_f$

Therefore we determine that $\rho^* = 0$.

The same reasoning can be applied if in equilibrium the license is signed between the lead imitator and the other imitator.

Therefore the equilibrium contract cannot involve a royalty rate. Furthermore, competition over the fixed fee will lead to an equilibrium contract $(\epsilon, 0)$.

The Financing Structure of Corporate R&D - Evidence from Regression Discontinuity Design

Nicolas Serrano-Velarde*
European University Institute

February 29, 2008

Abstract

I study the effect of a R&D subsidy on the investment and financing decisions of recipient firms. Locally unbiased identification is achieved using Regression Discontinuity Design. Empirical results are consistent with a stylized model of self-certification in which firm managers face imperfect information about the quality of their R&D. Firms for which R&D is an important part of their business model use the subsidy as a risk sharing device, substituting public funds for costly external finance. For "marginal innovators" the subsidy acts as a self-certification device that triggers an increase in private R&D investment driven by internal funding.

JEL classification: G32, H25, 031
Keywords: R&D Investment, Public Subsidies, Technology Policy, Regression Discontinuity Design, Policy Evaluation

*Nicolas Serrano-Velarde, European University Institute, Department of Economics, Via della Piazzuola 43, 50133 Florence, Italy, email: nicolas.serrano-velarde@eui.eu. I wish to thank Bruno Biais, Catherine Casamatta, Luigi Guiso, Andrea Ichino, Claire Lelarge, Paul Seabright, David Thesmar and Clifford Winston for their valuable comments. I also wish to thank members of the Committee of Statistical Secret and its Secretary Gérard Lang for providing the necessary data. All remaining errors are mine. The usual disclaimer applies

1 Introduction

A variety of instruments can be used by governments in order to overcome the gap between private and social returns to R&D: direct funding, tax credits, joint public-private research. However knowledge about the determinants of the financing structure of a firms' corporate R&D remains scarce. The issue is of importance since government intervention can only foster technological change if it supports projects that the private sector would not have implemented by itself.

This paper proposes and tests a stylized model of self-certification in which firms face imperfect information about the quality of their own R&D. The severity of this informational problem decreases with the R&D intensity and R&D investment of a firm.

The rationale behind this hypothesis is based on the observation that R&D activities are very information intensive and that managers' in firms have different collection costs of information. It results that in firms where R&D is just of marginal importance to the business strategy, managers will have less incentives to be informed about R&D projects and their quality. Consequently their beliefs about the probability of successfull R&D will be uncertain and lower than the actual probability of high quality research. Firms whose business model depends crucially on their ability to innovate will not face these internal doubts. If these firms want to do profitable business they have to innovate. Consequently managers of these firms have an incentive to be perfectly informed about the quality and the activity of their R&D department. In this setting, the subsidy acts as a certification device within firms that are "marginal innovators", but has no informational content for the "strong innovators".

The empirical test focuses on R&D subsidies given by the French ANVAR program, responsible for R&D support to small and medium sized firms. Since neither the firms receiving support, nor those not applying to a subsidy can be considered random draws identification strategies are required. Identification is achieved by using multiple eligibility rules in the subsidizing process of the ANVAR program. Regression Discontinuity Design (henceforth RDD) takes advantage of such discontinuous changes in the probability of receiving the subsidy as a function of some (pre)-observed covariate. In my case such discontinuous changes will be introduced through eligibility rules based on recipient firms' financial links. The identification strategy will allow me to obtain locally unbiased estimates of the effect of R&D subsidies on private R&D investment. Predictions from a stylized certification model closely match empirical results. Firms for which R&D is an important part of their business model use the subsidy as a risk sharing device, substituting public funds for costly external finance. For "marginal innovators" the subsidy acts as a self-certification device that triggers an increase in private R&D investment driven by internal funding. These results are not driven by potential scale effects in R&D activities

insofar as that for a given level of R&D investment the effect of the subsidy varies with the size of the firm in terms of employment. These results also reject the alternative hypothesis of financial constraints as driving the effect of the subsidy.

This paper attempts to adress the criticism voiced in David and Hall (2000) by combining a simple corporate finance model together with a quasi-experimental design. The quasi-experimental design is advantageous in this setting because, unlike other frameworks, it does not impose strong informational constraints. Empirical research on the complementary or substitutable nature of public support to R&D has indeed reached widely diverging conclusions. This is not astonishing insofar as these studies analyse different government programs, face different data constraints, and more importantly use different identification strategies.

Lichtenberg (1984, 1987) identifies endogeneity problems and measurement error as sources of bias in the estimation of the causal effect of subsidies. Lichtenberg (1987) shows how estimation can be affected in models which fail to control for shifts in the composition of final demand that are both correlated with federal R&D funding decisions and that determine R&D investment. I adress these issues by using a quasi-experimental approach in order to identify the effect of R&D subsidies on R&D effort of recipient firms. Both Duguet (2003) as well as Czarnitzki and Fier (2002) use matching methods in order to evaluate the impact of public R&D subsidies on private R&D investment. Duguet (2003) focuses on a sample of French firms from 1985 to 1997. He finds that, controlling for past public support, public funds add to private funds and that there is no significant crowding out effect. Although his study uses the same dataset as the present article, his study does not distinguish according to the source or the type of subsidy received. Similarly Czarnitzki and Fier (2002) study the effects of R&D subsidies on a cross-section of German service-sector firms. Although they reject the hypothesis of full crowding out they are only able to observe the participation status of firms. A similar data constraint is faced by Busom (2000). Using a cross-section sample of Spanish firms she finds that public funding induces more private effort but that for some firms full crowding out cannot be ruled out. Her results relate to the present paper insofar as it points at heterogeneity in the effects of subsidies across firms. The studies by Wallsten (2000) and Lach (2000) relate to my paper insofar as they focus on the allocation of subsidies by a specific agency, program. The study by Wallsten (2000) analyses the impact of the Small Business Innovation Act on a subset of publicly traded, young, technologically intensive firms in the US. Using a multi-equation IV model he finds that grants crowd out firm financed R&D spending on a $ by $ basis. Lach (2000) uses a dynamic panel approach in order to evaluate the effect of the Israeli OCS program on firm R&D. He finds that for small firms the subsidy

induces additional R&D investment, whereas for larger firms the subsidy has no statistically significant effect. A similar result is found by Gonzalez et al. (2005) using a model of firm R&D investment decision. They find that the increase in R&D investment due to subsidies is mainly driven by smaller firms. Finally, Klette and Moen (1998) point to the existence of positive dynamic effects of R&D subsidies using a panel data of Norwegian high tech firms. They find that subsidies increase private R&D after they expire consistent with learning by doing effects. An extensive survey of the literature can be found in David et al. (2000) as well as in Klette et al. (2000), who review some of the main findings of this strand of research.

2 Classical Theoretical Framework

I consider a classical firm investment model as presented in David et al. (2000).[1] At any point in time, an array of potential R&D investment projects is available to the firm. The firm is assumed to rationally consider expected costs and benefit streams for each project in order to calculate its expected rate of return thus forming its marginal rate of return schedule. Under the usual assumptions on returns to investment the MRR schedule will be downward sloping. The firm also has to consider the opportunity cost of investment. The marginal cost of capital schedule is increasing in the amount of R&D as increased R&D investment will be financed by external capital markets. The resulting equilibrium amount of R&D investment is denoted R*. More formally, using the notation of Klette and Moen (1998) one can represent the R&D investment of a firm receiving a R&D subsidy as:

$$R^* = R^{PP}(R^G) + R^{PG}(R^G) + R^G \tag{1}$$

Where R^* represents total R&D investment, R^G the subsidy received from the government, R^{PG} is the part of the subsidized project that is financed by the firm and R^{PP} is the R&D investment into non-subsidized projects. Privately funded R&D investment is a function of the subsidy. For simplicity I assume that firms receive matching grants, i.e. $R^{PG} = R^G$ [2], and that the total effect of the subsidy is measured by the effect of private investment

[1] Throughout the analysis I assume that the government agency responsible for allocating R&D subsidies wants to maximize total R&D (Duguet (2003)). Insofar as this is the only economically justified rationale for government intervention into private research this assumption seems plausible.

[2] It is straightforward to relax this assumption by defining the subsidiziation rate $t = \frac{R^G}{R^{PG}+R^G}$. Results will be qualitatively the same but with the presence of a multiplier $\frac{t}{1-t}$. For expositional convenience I present only the matching case.

into the subsidized project on the private investment into non-subsidized projects, i.e. $\frac{\partial R^{PP}}{\partial R^G} = 0$. Since I am interested in the net effect of the subsidy on private R&D investment I define the net of subsidy private investment as $R^*_N = R^* - R^G$. The full effect of the subsidy on net private R&D investment is now given by

$$\frac{dR^*_N}{dR^G} = \frac{\partial R^{PP}}{\partial R^G} + \frac{\partial R^{PP}}{\partial R^{PG}} \cdot \frac{\partial R^{PG}}{\partial R^G} + \frac{\partial R^{PG}}{\partial R^G} + \frac{\partial R^{PG}}{\partial R^{PP}} \cdot \frac{\partial R^{PP}}{\partial R^G} \tag{2}$$

Which simplifies to

$$\frac{dR^*_N}{dR^G} = \frac{\partial R^{PP}}{\partial R^{PG}} + 1 \tag{3}$$

From equation 3 it clearly appears that subsidies spur private R&D investment only if the subsidized projects would not have been implemented in the absence of the subsidy, i.e. $\frac{\partial R^{PP}}{\partial R^{PG}} \geq 0$. The framework derived in the preceding section can be used to discuss the effect of R&D subsidies on private R&D investment. For each case I review which type of economic constraints could generate these effects.

Benchmark Case, Complementary Effect: $\frac{dR^*_N}{dR^G} = 1$ and $\frac{dR^{PP}}{dR^{PG}} = 0$. The subsidy increases firm total R&D investment by 2 Euros: 1E of subsidy and an additional 1E of privately financed R&D investment, leaving unaffected the investment on non-subsidized projects. In case of perfect information the agency is able to perfectly identify the marginal project such that private funding of non-subsidized projects is unaffected. Conceptually such a complementary effect of the subsidy can also arise in the presence of imperfect information. I distinguish 2 types of constraints that could generate such an outcome.

1. Credit Constrains (Guiso (1997)): Firms are credit constrained and are not able to finance their privately optimal level of R&D investment. The subsidy relaxes the constraint and enables the firm to get banks on board. In this situation cash responsiveness of R&D investment, especially for high-tech firms, is most likely due to pervasive credit constraints rather than to cash flow proxying for future expectations.

2. Technological Complementarities: Captures the idea that the more a firm invests into R&D the more profitable this investment becomes. In a dynamic setting such complementarities are reflected by a learning by doing process (Klette and Moen (1998)).

- Spillover effects (Lichtenberg (1987)): Knowledge that increases the productivity of privately employed R&D inputs, thus lowering its private costs.

- Joint Cost Structure of R&D (Duguet(2003)): the cost structure of R&D activities is such that investing into the necessary fixed sunk costs of the subsidized project might turn additional projects profitable for the firm. The joint cost structure of R&D activities gives rise to potential scale effects that I will discuss later.

Substitution Case: $\frac{dR_N^*}{dR^G} = -1$ **and** $\frac{dR^{PP}}{dR^{PG}} = -2$**.** The subsidy reduces total R&D investment by 1E: the firm receives 1E of subsidy which it matches with 1E of private funding but at the same time reduces investment into non-subsidized projects by 2E. Conceptually such a substitution effect could derive from 2 situations.

1. Portfolio reorganization in the context of inelastic supply of R&D inputs (Lach (2000), David et al. (2000), Lichtenberg (1984)): the subsidy turns an unprofitable project into a profitable one but hiring cost of R&D personel are such that the firm has to discontinue another profitable project (assuming the firm committed to implement the subsidized project).

2. Subsidization of profitable projects: The government subsidizes projects which would have been implemented even without the subsidy. However one can distinguish again 2 possible sources of direct substitution.

- Asymetric information (Klette and Moen (1998), Duguet (2003)): The government agency is not perfectly informed about the firms' optimal level of R&D without subsidies and since public funds are cheaper than private funds, firms apply for projects with sufficiently high returns.

- Winner Picking (Wallsten (2000)): The government agency is under pressure to pick projects that are most likely to succeed. The winner picking story can either arise because of political pressure on decision-makers eager to avoid negative publicity for their projects, or because the agency finances itself from repayment of subsidized loans. In both cases the government agency is willing to subsidize projects which would have been implemented in the absence of the subsidy.

3 A Stylized Model of Self-Certification

The proposed model of self-certification is based on the idea that, in addition to technological uncertainty, there is also managerial uncertainty about the quality of the firms' R&D activities. The severity of this informational problem decreases with the R&D intensity and investment of a firm. The rationale behind this hypothesis is to model a reduced form of the organizational aspect of firm investment. In firms where R&D is just of marginal

importance to the business strategy, managers will have less incentives to be informed about R&D projects and their quality. Consequently their perception about the quality of their R&D will be lower than the actual probability of successful research. Firms whose business model depends crucially on their ability to innovate will not face these internal doubts. If these firms want to do profitable business they have to innovate. Consequently managers of these types of firms have an incentive to be perfectly informed about the quality and the activity of their R&D department.

Setup of the Model

The agents

There are 2 types of firms in the economy which have different information about the quality of their R&D according to the importance of R&D in their everyday business. The high type firms perfectly perceive their probability of successful or unsuccessful R&D $P_H = 1$. Low type firms on the other hand face imperfect information about the quality of their R&D activities. They imperfectly observe the "'true"' probability of successful R&D, i.e. $P_L < 1$. The distinction between types is based on the importance of R&D activities in a firms' business model. For high type firms R&D is a crucial aspect of their business strategy and therefore managers have an incentive to be informed about the quality of their R&D Department. They constitute the "strong innovators". The business strategy of low type firms is only marginally based on their ability to innovate and therefore managers are less likely to be informed about the quality of their research. They constitute therefore the group of "marginal innovators".

Investment, Financing and Returns to Innovation

A firm can invest an amount R into R&D activities, $R \in [0, R*]$. In addition to a managers' uncertainty about the general quality of R&D in his firm there is also intrinsic uncertainty about the success of a project. Consequently R&D can have a rate of return of either K or O with probability p and $(1-p)$

$$= \begin{cases} K & \text{with probability} \quad p \\ 0 & \text{with probability} \quad (1-p) \end{cases}$$

Now, every firm can finance its R&D investment with initial cash A, $A \in]0, R*[$, at an intertemporal rate of discount ρ. The firm can also search for external funding at additional cost of λ. The payoff structure will depend on the information set of the manager of the firm. If the firm has perfect information and it is sure about the intrinsic quality of the project, then: $P_H.p.K \succ (1+\rho+\lambda)$, i.e. it will be able to pay external

costs λ. However, if the firm has imperfect information about the quality of its research, then it will not be able to see whether one or even all the projects are of good quality. Consequently, they will not invest if their uncertainty about the probability of succesfull research is low enough $P_L.p.K \prec (1+\rho)$.

Subsidy

Firms can obtain an R&D subsidy from a public agency. The behavior of the agency is not explicitly modeled, however I assume it has a signaling value with respect to the quality of the firms' R&D department. This assumption seems plausible insofar as most public agencies have a technical review process of submitted projects. Public financing is assumed less costly than external financing, which allows some generality with respect to the different forms of subsidies [3]. No additional assumption is needed on the side of the agency. Following reception of the subsidy, firms update positively their beliefs about the quality of their own R&D investments.

Investment Decisions in the Absence of a Subsidy

Since managers of "strong innovators" perfectly perceive the true probability of successful R&D, $(P_H.p)$, they will invest:

$$R^* = A + B \tag{4}$$

It will then receive and share the following payoffs:

$$p.K.R^* = \begin{cases} (1+\rho+\lambda)B & \text{for the bank} \\ \pi \succ 0 & \text{for the firm} \end{cases}$$

Consequently the "strong innovator" invests the optimal amount into R&D. The fact that the manager knows the intrinsic quality of its R&D investment allows it to raise costly external finance.

Without the subsidy the "marginal innovator" has an expected rate of return from R&D which is $P_L.p.K$. Provided managers are sceptic enough about the quality of their firms's R&D, they do not invest into R&D (neither private nor external).

$$P_L.p.K \prec (1+\rho) \tag{5}$$

Investment Decisions with a Subsidy

[3] It does not matter whether it is a pure subsidy or only a re-imbursable loan at advantageous rates as long as these funds are less costly than private ones

Since the subsidy carries no new information for managers of "strong innovators" total R&D investment will not change. However the amount of external finance changes since it is assumed more costly than public finance, the subsidy will be used to reimburse part of B to avoid cost λ. Hence the subsidy is used as a risk sharing device.
Upon reception of the subsidy managers of "marginal innovators" update their beliefs about the quality of their R&D from P_L to $P_L^{UPDATED}$ with $P_L^{UPDATED} \succ P_L$. Consequently it expects a high rate of return K with probability $P_L^{UPDATED}.p \succ P_L.p$. Provided the signal is clear enough, the firm will start investing own funds but not external funds if

$$1 + \rho + \lambda \succ P_L^{UPDATED}.p \succ 1 + \rho \tag{6}$$

The firm will also start investing external funds provided that

$$1 + \rho + \lambda \prec P_L^{UPDATED}.p \tag{7}$$

If the firm starts investing into R&D but does not access external funds its investment into R&D will be $r = S + A \leq R^*$. This means that even if the "marginal innovator" receives a subsidy it will invest a smaller amount into R&D than in the case of perfect information.

Empirical Testing of the Self Certification Hypothesis

Summary of Predictions
In the absence of subsidies "strong innovators" invest their privately optimal amount of R&D R^*, taking into account the technological (intrinsic) uncertainty, and complement their internal financing with costly external funding B. "Marginal innovators" not only face technological uncertainty but also managerial uncertainty and therefore refrain from investing internal or external funds into R&D.
Upon reception of the subsidy high type firms do not modify their privately optimal amount of R&D R^*, they simply substitute external funding with public funding. Low type firms on the other hand update their beliefs about the quality of their research and start investing internal funds into R&D. Depending on the strength of the signal, these firms may even start accessing external funding.

Empirical Estimation
The self-certification model predicts a differential effect of the subsidy according to the importance of R&D in the business strategy of the firm. The testing procedure will acommodate this heterogeneity by using quantile regression methods. The effect of the subsidy

on private R&D investment will thus vary according to the absolute amount of private R&D investment. Provided absolute R&D investment is correlated with the importance of R&D within a firm (controlling for size, industry and other controls) the testing procedure approximates well our distinction according to types[4].

Potential Caveats

One potential caveat is however to confuse the self-certification hypothesis with a simple scale effect in R&D activities. In other words, one might be worried that the differential effect in subsidies arises because the subsidy enables the firm simply to cross a critical threshold in terms of R&D scale. This concern can be adressed using the differential predictions of the models with respect to the non-R&D size of the firm. Indeed if the effect of the subsidy is simply related to the size of the R&D budget then the size of the firm in terms of non-R&D employment should not matter. This is not true anymore in the case of the self-certification hypothesis. If self-certification is driving the differential effect of the R&D subsidy, then for a given level of R&D investment the signal of the subsidy will be stronger in larger firms in terms of non-R&D employment. Another advantage of this testing strategy is that it also allows to discriminate the self-certification hypothesis from a financial constraints hypothesis. If firm size can be taken as a proxy for financial constraints then the financial constraint hypothesis can be rejected against the self-certification hypothesis if (i) firms at lower and higher quantiles of R&D investment are not statistically different in terms of size, (ii) the effect of the subsidy is lower the larger the firms (i.e. the less likely financial constraints). Consequently the differential implications of the self-certification model with respect to size of the firm can be used to discriminate against potential scale effects and/or the financial constraint hypothesis.

4 Stylized Facts and Institutional Framework

The ANVAR Program

ANVAR was created in 1979 to support R&D projects of small and medium sized firms through reimbursable aid[5]. Every year Anvar supports between 1.000 and 1.500 projects for a total budget of 250 M Euros. Aid is paid on the basis of advancement of the project. Projects are selected on the basis of a bottom-up process by which firms

[4]In addition it can be argued that, controlling again for other characteristics, the total R&D investment proxies well the types. Indeed, everything else constant, firms which feel more confident about the quality of their research will engage into more R&D

[5]I will use reimbursable aid, public support and subsidy interchangeably in this context

propose their projects to the agency. The agency has no specific mandate on what type of R&D or sector they can fund. Rejection rates are low and once the agency grants its support to a project it also helps the firm to find potential private/public partners.

The empirical analysis combines the yearly R&D Survey from the Ministry of Research and the Financial Links Survey from INSEE. Table 1 reports the descriptive statistics relative to the sample. Pooling the data over the 1995-2004 period the database amounts to 21087 firm-year observations. I distinguish between firms in the overall sample and firms that received a subsidy. Approximately 11% of the firms in the sample received financing from ANVAR. Subsidized firms were on average significantly smaller than firms in the full sample in terms of both total sales and employment. These differences seem less pronounced in terms of innovation characteristics, especially with respect to the number of researchers employed. This suggest that ANVAR financed firms are less commercially developped and more research oriented. The financing by ANVAR amounts up to 8 MF and ranges from 100KF at the 25th percentile of the distribution to 850 KF at the 75th percentile. Although in our sample the total amount of financing from ANVAR increases over the period, accounts from ANVAR suggest that the total budget was slowly decreasing until 2004.

Table 1: Descriptive Statistics for Sample and Treated Firms

			Percentiles			
	Mean	St.Dev.	p25	p50	p75	Number of Observations
Total Sales	1469	9568	59	214	678	21087
Total Sales for Subsidized Firms	311	3483	10	49	163	2312
Employment	1132	8639	73	222	602	21087
Employment for Subsidized Firms	287	3266	24	70	188	2312
Net of Subsidy R&D	61.5	359	1.9	5.5	20.2	21087
Net of Subsidy R&D for Subsidized Firms	20	179	1	3	9	2312
Number of Researchers	32	164	2	5	14	21087
Number of Researchers for Subsidized Firms	13	70	2	4	10	2312
Number of Patents*	27.8	139	2	4	13	4576
Number of Patents for Subsidized Firms*	7.24	17.8	1	3	6	664
Subsidy*	0.695	0.943	0.111	0.375	0.845	2312

*Only for Variable¿0
Total Sales, Net of Subsidy R&D Investment and Subsidy Variables expressed in milions of Francs (1995).

Eligibility to the Program

In order for a firm to be eligible to the ANVAR program it has to have less than 2000 employees and to be independent from a large business group (henceforth referred to as

Table 2: Full Sample and RDD Sample

	(Marginally) Eligible		(Marginally) Ineligible		Total
	Treated	Non-Treated	Treated	Non-Treated	
Full Sample	2156	13073	156	5702	21087
(1995-2004)	10.2%	62%	0.7%	27.1%	100%
Total	15229 Eligible		5858 Ineligible		21087
	72%		28%		100%
RDD Sample	86	294	0	186	566
(1995-2004)	15.2%	52%	-	32.8%	100%
Total	380 Eligible		186 Ineligible		566
	67%		33%		100%

LBG). Independence is defined with respect to the firms' ownership structure. According to French law a firm is independent if less than 25% of its capital is owned by a LBG[6]. Although the agency has a certain discretion with respect to the definition of independence I use the legal definition of independence. The data shows that this is indeed the relevant threshold considered by firms when applying. Consequently, a firm owned at 50% by a Business Group of 1000 employees will be considered eligible in this setting. A firm owned at 26% by a Business Group of 2001 employees will be considered ineligible in this setting. For each firm I first identify shareholders and their respective shares in the capital of the firm. To each shareholder I assign the total employment of the group it represents. The identification strategy being only valid at the threshold of eligibility I restrict the sample to firms which have $0\% < X < 50\%$ ownership by a LBG. Consequently, in the RDD sample a firm is marginally-eligible whenever it has positive ownership by a LBG but its share does not exceed 25%. Secondly a firm is marginally-ineligible whenever it has between 25% and 50% of its capital owned by a LBG. I further restrict eligibility conditions to be binding only when LBG consist of industrial partners. I do so by not taking into account ownership by public firms, banks and national champions. The first two categories are easily identified in the data. On average I however have 2 formally ineligible firms per year that obtain treatment. I also exclude them from the sample because their LBG consists either of agricultural cooperatives or of public related firms which escape firm

[6]JO L 107, 30.4.1996, p. 4 : "'are considered independent firms whose capital or voting rights are not owned more than 25% by firms not corresponding to SME classification"'.

categorization. The 1995-2004 full sample consists of 22000 firm observations, whereas the 1995-2004 RDD sample consists of 500 firm observations. Within the RDD analysis I distinguish between four bandwidths around the threshold : Large ($0\% < X < 50\%$), Intermediate ($5\% < X < 45\%$), Small ($10\% < X < 40\%$), Very Small ($15\% < X < 35\%$). The smaller the bandwidth, the more likely are the conditions of a quasi-experiment. However one has to keep in mind the trade-off between length of the bandwidth and number of observations. In the analysis I restrict interpretation of the results to the Small and Very Small bandwidth and use the larger bandwidths mainly to check robustness.

5 Identification Strategy

Dependent Variable. In the empirical analysis I will first use private investment into R&D as the relevant outcome variable. According to the stylized model of self-certification the subsidy should have a complementary effect at lower quantiles of the R&D distribution and no effect at higher quantiles of the R&D distribution. The analysis will then decompose R&D investment into its internally and externally financed components. Again, according to the self-certification hypothesis the complementary effect at lower quantiles of the distribution should be driven by increased internal financing whereas public financing should be substituted for external financing at higher quantiles.

The Causality Problem and RDD framework. Results based on the assumption of unconfoundedness are dubious since neither the firms receiving support, nor those not applying to the ANVAR program can be considered random draws. Because of this endogeneity problem OLS is unlikely to estimate an unbiased causal effect of the subsidy[7]. RDD is a quasi-experimental design taking advantage of discontinuous changes in the probability of receiving treatment as a function of some (pre)-observed covariate. Assignment to treatment solely depends on whether pre-intervention variables satisfy one or a set of conditions. Let Y represent the outcome of interest, i.e. the net of subsidy R&D investment by firms. Let S, the assignment covariate, be the maximum ownership by a LBG with known eligibility threshold $\bar{s}$ at the 25% ownership level. Finally denote by T the actual treatment status of the firm. Identification in a RDD setting is defined if

$$(Pr[T = 1 \left| \bar{s}^{+} \right.] \neq Pr[T = 1 \left| \bar{s}^{-} \right.]) \tag{8}$$

Where $\bar{s}^{+}$ and $\bar{s}^{-}$ refer to those firms marginally above and below the threshold. Depending on the size of the discontinuity one obtains either a sharp or a fuzzy design.

[7]Plain OLS regressions suggest a statistically significant effect of the subsidy with implausible magnitude

Sharp RDD occurs if the probability of treatment increases from 0 to 1 as S crosses the threshold of $\bar{s}$, whereas in the case of fuzzy design the jump is smaller than one. However the analytical framework of this study departs from the usual RDD framework in an important institutional dimension. Ineligible firms in this setting have a 0 probability of receiving the subsidy, whereas eligible firms have a positive non-deterministic assignment to treatment. Following Battistin and Rettore (2007) the regularity conditions required to achieve identification in this setting are the same as in the sharp setting. Consequently the following condition is sufficient to identify the mean impact of the subsidy at $\bar{s}^{+}$:

Condition The mean value of Y_0 conditional on S is a continuous function of S at $\bar{s}$

An interesting interpretation of this condition in terms of unobserved types can be found in Lee(2001). At the threshold for eligibility only the probability of treatment changes discontinuously. Intuitively, in a neighborhood of $\bar{s}$ RDD presents the same features as a pure randomized experiment. Exploiting the relationship between S and T

$$(Y_1, Y_0) \perp T \left(S| = \bar{s}\right) \tag{9}$$

An appealing feature of RDD is that underlying assumptions can be tested. Under random assignement, any variable determined prior to the assignment will have the same distribution in the treatment and control groups. Threats to RDD identification will be tested in section 4. Exploiting the regularity condition and taking into account the non-deterministic nature of assignment to treatment on one side of the assignment covariate (ineligible firms have a 0 probability of treatment but eligible firms do not all benefit from treatment) Battistin and Rettore(2007) show that:

$$E\left[y_{i0}|\left(S = \bar{s}^{+}\right)\right] = E\left[y_{i0}|\left(S = \bar{s}^{-}\right)\right] \tag{10}$$

The LHS can be written as a weighted average of the mean outcome for eligible participants and the mean outcome for eligible non-participants,

$$E\left[y_{i0}|\left(T_i = 1, S = \bar{s}^{+}\right)\right] \cdot \phi + E\left[y_{i0}|\left(S = \bar{s}^{+}\right)\right] \cdot (1 - \phi) = E\left[y_{i0}|\left(S = \bar{s}^{-}\right)\right] \tag{11}$$

Where $\phi = E[T|S = \bar{s}^{+}]$.This implies that the counterfactual mean outcome for marginal participants is identified by a linear combination of factual mean outcomes for marginal ineligibles and marginal eligibles not participating into the program no matter how participants self-select. The estimation strategy will use this result in order to disaggregate the effect of the subsidy at different levels of private R&D investment. Building on this

intuition I estimate directly the effect of the subsidy around the threshold through OLS and quantile regression techniques. The estimated coefficient on the treatment will incorporate variation between treated units and variation between treated and non-treated units. According to equation 11 these are valid counterfactuals in the 1 Sided Fuzzy Design. In this estimation framework the weights differ from those in equation 11 but I can test their importance by estimating the equations on different bandwidths. Varying the bandwidth changes the proportions of non-treated eligible and ineligible firms. If estimates are robust with respect to such variations the weights are unlikely to play a crucial role in this setting.

6 Empirical Analysis

Relevance of the RDD Framework

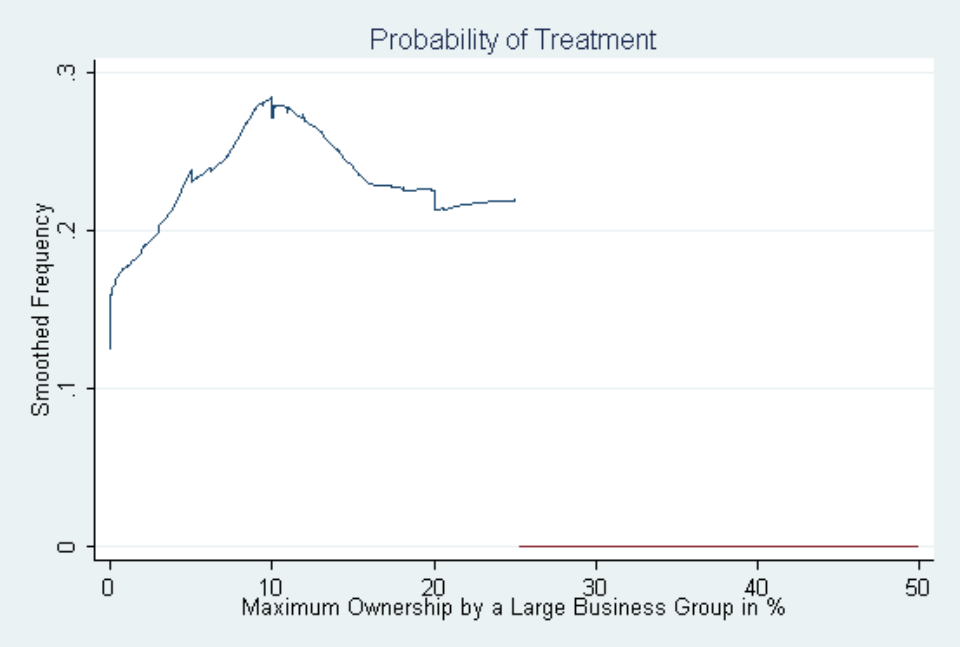


Figure 1: Probability of Treatment as a Function of LBG Ownership

Figure 1 presents graphical evidence for the existence of a discontinuous change in the probability of receiving treatment as a function of the ownership variable. Figure 1 estimates a locally weighted smoothing regression separately above and below the threshold of eligibility. A jump in the plot shows the effect of the threshold on the probability of receiving financing from ANVAR. It shows that there is a substantial effect of the eligibility threshold on the probability of the firm to receive treatment. Figure 2 reproduces the analysis of Figure 1 using the net R&D investment by firms as a dependent variable. At

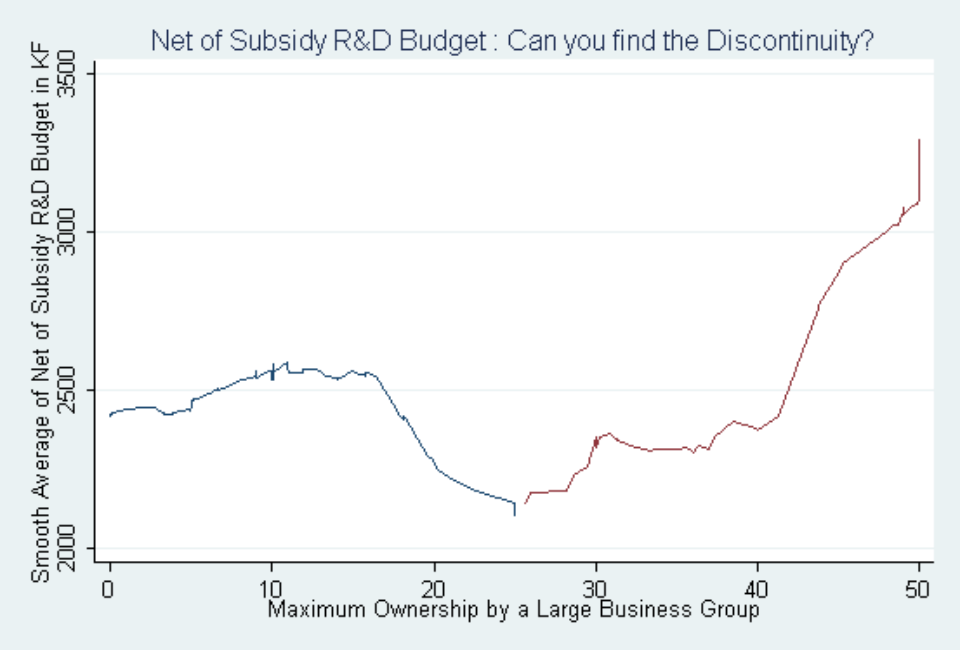


Figure 2: Net of Subsidy R&D Investment as a Function of LBG Ownership

the threshold for eligibility I do not observe a significant jump in the net R&D investment by firms.

Testing for Self-Selection into Eligibility

A challenge to the RDD framework is potential self-selection into eligibility status. According to Battistin and Rettore (2007) if some ineligible subjects alter their ownership structure so as to switch from values above the threshold to values below it, a discontinuity in the cumulative distribution of the observed ownership pattern will be found at the threshold. Figure 3 shows that the distribution of firms' ownership displays discontinuities at the 0% and the 50% and 100% thresholds. At the threshold for eligibility there does not seem to appear a discontinuity, i.e. firms do not cluster either above or below the 25% ownership threshold. Graphical evidence is thus consistent with theoretical and legal evidence. On theoretical grounds it is not clear that ownership structures can be understood as a choice variable. The corporate finance literature has shown that ownership decisions may depend on interactions and coalitions between shareholders as in Bennedsen and Wolfenzon (2000). A firms' initial owner in need of external capital sells votes and cash-flow to outside investors. Insofar as control induces private benefits the initial owner will take into account both the need to raise funds and the contest for control within the firm. Depending on which effect dominates different coalitions and different ownership structures may arise. Consequently an outside investor might not have a direct

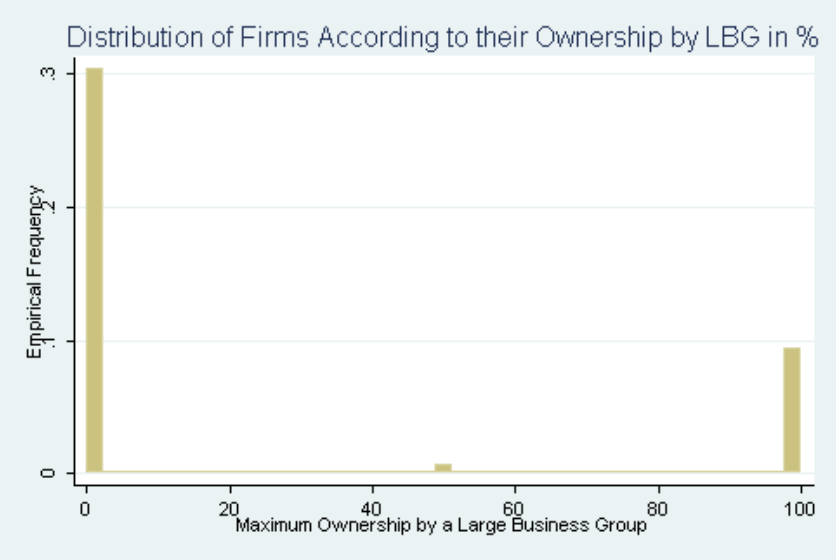


Figure 3: Distribution of Firms as a Function of LBG Ownership

control of his stake in the firm [8]. In addition, casual evidence suggests that the fiscal and legal environment in France does not vary greatly according to either ownership or SME status. Indeed French firms pay a flat tax rate (33%), and labor laws are a function of employment only.

Testing for Validity of the Continuity Condition

To gather evidence on the validity of the continuity conditions on which the identification strategy relies I implement an overidentification test following Lee (2006) and Giavazzi et al (2007). Consider the set of pre-intervention variables verifying two requirements : they are stochastically related to the base outcome but logically unaffected by the program. Then I check whether the groups above and below the threshold are balanced with respect to such a variable. If firms around the threshold differ with respect to these variables, the identification strategy fails since they are no longer comparable in terms of relevant unobserved components. I take as pre-intervention variables R&D information plausibly related to the unobserved component but that could not have been modified by the subsidy. The variables considered are locational choice, whether or not a firm received alternative subsidies, whether or not the firm is classified high-tech, whether or not the firm undertakes process and product innovation. As far as the propensity to do product and process innovation is concerned I need an additional assumption to ensure that firms did not modify their R&D strategy so as to obtain the subsidy. The assumption seems plausible insofar as the program is based on a bottom-up approach for participation and

[8] at least not in increasing its share in a firm

Table 3: Tests for the Presence of Sorting and for the Validity of the Continuity Condition

	Bandwidth			
	(1) Large ($0\% < X < 50\%$)	(2) Intermediate ($5\% < X < 45\%$)	(3) Small ($10\% < X < 40\%$)	(4) Very Small ($15\% < X < 35\%$)
Process Innovation	-0.55	-.148	-.156	-.044
	(.127)	(.148)	(.175)	(.23)
Product Innovation	.026	.01	-.168	-.026
	(.139)	(.155)	(.2)	(.232)
Alternative Subsidies	.079	.041	.144	.202
	(.131)	(.141)	(.180)	(.211)
Location "Paris"	-.239	-.112	-.179	.285
	(.134)	(.141)	(.181)	(.216)
High-Tech Firms	-.012	-.023	.03	.117
	(.088)	(.096)	(.116)	(.150)
Time Effects	Yes	Yes	Yes	Yes
Industry Effects	Yes	Yes	Yes	Yes
Number of Observations	560	380	276	189

Each coefficient (and related robust-clustered standard error in parenthesis) is an estimate
of the coefficient on eligibility obtained from separate regressions of the form :
S=h(LBG) + (E)β + Year + Industry
Where S is the pre-intervention outcome indicated in the corresponding row of the table; Eligibility is a dummy indicating whether a firm is eligible or not.
Estimates are obtained using OLS. Year stands for time dummies. Industry stands for Industry Fixed Effects.

Table 4: Mann-Whitney Test for the Validity of the Continuity Condition

	Fundamental Research	Applied Research	Experimental Research
Mann-Whitney	.324	1.58	-.741
	(.74)	(.113)	(.47)
H0	Not Rejected	Not Rejected	Not Rejected

Eligibility as sorting criterion; p-values between brackets

since the program has no specific incentive schemes. In addition, if the program modifies R&D of the eligible firms then this only biases our test estimates in favour of self-selection around the threshold.

The test is implemented by running a regression using as a dependent variable the battery of pre-intervention outcomes (P) and eligibility status as an explanatory variable (E).

$$P = \beta_0 + E \cdot \beta 1 + h(LBG) + Year + Industry \tag{12}$$

I also include a polynomial in ownership as well as year and industry fixed effects[9]. The estimates indicate that eligibility to the program is associated with a decrease of 1.2% points in the probability of being a R&D intensive firm. This estimate is small, statistically not different from zero and its sign is opposite to the one expected under the sorting hypothesis. Similarly insignificant is the estimate with respect to alternative sources of financing or product innovation. If eligible firms were more likely to be say financially constrained than ineligible firms, I would expect eligible firms to more systematically resort to public financing. We can, therefore, exclude the existence of sorting around the thresholds on the basis of eligibility. The rest of Table 3 presents evidence on other pre-intervention outcomes that should not be affected by eligibility status while depending on the same unobservables (i.e. creativity or financial constraints), likely to affect private R&D investment. Again, coefficients are small in size and statistically non-significant. Locational choice for instance is statistically insignificant and its sign is opposite of the one expected under sorting, i.e. more creative firms clustering around Paris. These results are unlikely to be driven by sample size insofar as results are robust to the extension of the bandwith. Finally Table 4 uses a rank-sum test to test whether the different type of R&D

[9] All results are robust to the inclusion of additional explanatory variables

distributions above and below the threshold differ significantly. The test does not reject the null hypothesis that firm expenditures into fundamental, applied and experimental research is similar across the two groups.

7 Results

Quantile Estimates: Static and Dynamic Effects

The following specification will constitute the basis for the empirical investigation:

$$Y = \beta_0 + S \cdot \beta 1 + X \cdot \beta 2 + h(LBG) + Year + Industry \tag{13}$$

where h(LBG) is a second order polynomial in LBG. Employment, Total Sales and R&D intensity are included so as to capture potential size effects and thus reduce the amount of heterogeneity in the subsidy. I include year-specific effects in order to account for potential time variation of the variables. Finally I include industry specific effects. In my final and reported specification I exclude the polynomial in ownership for theoretical and practical reasons. Practically, estimates never appear indiviually and jointly statistically different from 0 at all conventional levels. Secondly the control function is meant to capture the effect of the assignment covariate on the dependent variable as estimation includes larger and larger bandwidths. However there is no theoretical argument on how ownership by a large business group should affect private R&D investment of a firm, especially in the case of non-controlling shareholders. Estimates are reported for four considered bandwidths. Pursuing the argument made in the theoretical section I directly ask the question whether subsidies affect total firm investment in the same way using quantile regression. The direct use of the subsidy in is motivated by the special 1 Sided Fuzzy Design of this study. Indeed following Battistin and Rettore (2007) in this special setting eligible non-treated and ineligible firms are valid counterfactuals for supported firms, no matter how these supported firms self-select into the program.

Results from quantile regression are not only statistically significant and robust, but are also economically sensible. For all bandwidths considered I find a statistically significant positive effect of the R&D subsidy on private R&D investment for firms at the lowest quartile of the private R&D investment distribution. For firms with relatively smaller R&D budgets, an additional Franc of subsidy increases their own R&D investment by 1.1 Francs. This result is consistent with a Franc by Franc matching of the subsidy on the side of the recipient firm [10]. Interestingly median regression reproduces the results from

[10]The coefficient on subsidies being higher than 1 does not suggest a stronger complementary effect

OLS and First Differences estimation insofar as the subsidy is at best additive for these firms. A result maintained at higher quantiles of the distribution. This result is consistent with evidence found by Lach (2000) and to a certain extent by Busom (2000). In Lach (2000) the subsidy given by the Israeli OCS program has a strong complementary effect for small firms but no statistically significant effect for large firms. In Busom (2000), R&D subsidies given by Spanish authorities have a heterogeneous effect on firm R&D: in 30% of the cases the subsidy leads to crowding out effects, whereas for the rest of the firms the subsidy induces more private effort. The robustness of the results with respect to bandwidth specification makes me confident in the validity of the estimation framework. Indeed, according to equation 11 the validity of the quantile regression estimates depends on the optimal linear combination of eligible non-participants and ineligible firms. If the quantile estimates were to vary greatly according to the bandwidth it would suggest that changing the proportions of non-treated eligibles and ineligibles plays a crucial role in the estimation. In addition, the weights attached to non-treated eligibles and ineligibles are important only when their base outcome performances differ greatly. As this does not seem to be the case judging from two sample t tests (cf. Appendix) the interpretation of the results as being quantile treatment effects seems warranted. A further re-assuring feature of the results is their robustness with respect to the inclusion/exclusion of additional covariates. Indeed, as mentionned in Lee (2007), if the hypothesis of a randomized experiment is valid then results should not vary greatly with inclusion/exclusion of covariates (only precision might vary).

The next step in the analysis is to investigate possible dynamic effects of the subsidy. Such effects could arise in the presence of learning by doing effects as found in Klette and Moen (1998). I re-estimate introducing the lag of the subsidy in addition to the actual subsidy received during the year. Unfortunately, due to the survey method used by the Ministry of Research there is a high correlation between receiving a subsidy / not receiving a subsidy and being questionned for the survey [11]. Interpretation of the results should bear in mind these limitations. The results are globally inconsistent with the hypothesis of dynamic effects of the subsidy. The coefficient on the lagged subsidy variable is not statistically significant in most cases. The estimates of the contemporaneous effect of the subsidy remain unchanged with respect to the inclusion of the lagged subsidies variable. It is possible that the results are driven by the low power of the test, however if this is not the case my results suggest that the programs' support does not change the technological path of the firm. The subsidy allows firms at the lowest quartile of the R&D distribution to implement a given project that otherwise would not have been implemented, however

but could simply be due to the departure from the matching assumption: public support covers generally between 30% and 50% of the projects costs

[11] Since 1998 the Ministry of R&D systematically surveys firms receiving subsidies from ANVAR.

it does not lead them to change their optimal R&D strategy. In the terminology of Lach and Sauer (2002), receiving a subsidy increases the likelihood of implementation / development of the project, but does not affect research efforts.

Table 5: Quantile Regression Estimates

	Quantile Regression Estimates			
	.25 Quantile	.5 Quantile	.75 Quantile	Observations
Large Bandwidth	1.12	-.13	1.13	560
$(0\% < X < 50\%)$	(.29)	(1.16)	(2.01)	
IntermediateBandwidth	1.17	-1.55	2.55	380
$(5\% < X < 45\%)$	(.32)	(2.4)	(5.26)	
Small Bandwidth	1.10	-.84	1.78	276
$(10\% < X < 40\%)$	(.23)	(1.5)	(4.93)	
Very Small Bandwidth	1.33	-2.7	-4	189
$(15\% < X < 35\%)$	(.40)	(1.6)	(1.21)	

Each coefficient (and related standard error in parenthesis) is an estimate of the coefficient on the subsidy obtained from separate regressions of the form :

$$Y = \beta_0 + S \cdot \beta 1 + X \cdot \beta 2 + h(LBG) + Year + Industry$$

Where S is the actual level of treatment indicated in the corresponding row of the table;
X is a vector of covariates.
Estimates are obtained using quantile regression.
Year consists of Time Dummies. Industry consists of Industry Fixed Effects.

Table 6: Quantile Regression Estimates - Dynamic Effects

		Quantile Regression Estimates			
		.25 Quantile	.5 Quantile	.75 Quantile	Observations
Large Bandwidth	Static	1.17	.38	2.5	560
$(0\% < X < 50\%)$		(.32)	(1.3)	(2)	
	Dynamic	-.18	-1.2	-4.7	
		(.4)	(1.03)	(1.5)	
IntermediateBandwidth	Static	1.22	-1.56	1.68	380
$(5\% < X < 45\%)$		(.37)	(1.76)	(7)	
	Dynamic	-.4	-1.52	-6.4	
		(.56)	(1.81)	(5.1)	
Small Bandwidth	Static	1.10	-2.38	3.32	276
$(10\% < X < 40\%)$		(.18)	(1.54)	(4.68)	
	Dynamic	-.11	-2.58	-5.17	
		(.3)	(1.63)	(3.76)	
Very Small Bandwidth	Static	1.24	-2.5	-4	189
$(15\% < X < 35\%)$		(.35)	(1.91)	(.81)	
	Dynamic	1.07	-2.67	-3.4	
		(.31)	(1.89)	(.82)	

Each coefficient (and related standard error in parenthesis) is an estimate of the coefficient on the subsidy obtained from separate regressions of the form :

$$\text{Y}= \beta_0 + \text{(Subsidy)}\beta_1 + (Subsidy_{-1})\beta_2 + \text{X}\cdot\beta 3 + \text{h(LBG)} + + \text{Year} + \text{Industry}$$

Where Subsidy, denoted Static, is the actual level of treatment indicated in the corresponding row of the table;
$Subsidy_{-1}$, denoted Dynamic, is the lagged value of the subsidy;
X is a vector of covariates.
Estimates are obtained using quantile regression.
Year consists of Time Dummies. Industry consists of Industry Fixed Effects.

Quantile Estimates: The Financing Structure of Corporate R&D

I further explore the impact of the subsidy on the financing structure of corporate R&D. With the dataset at hand I am able to decompose private R&D investment into internally and externally financed R&D investment. Although in my data externally financed R&D includes not only bank loans but also payments from other firms, this measure can be interpreted as a proxy for banking finance [12]. Finally I construct a measure of external financing dependence defined as the ratio of externally financed R&D with respect to total private R&D. I re-estimate the effect of R&D subsidies on each of these components in order to gain a better understanding of how the financing structure of the firm evolves. The fact that a non-negligible fraction of firms reports no external finance can however make quantile estimation more difficult. I check robustness by first estimating separately equations for firms with positive values of external finance and then use censored quantile regression methods. Censored quantile regression is an iterative process based on Buchinsky (1991,1994) in which one first estimates an unrestricted quantile regression and then excludes observations at a given quantile with predicted values below the censoring points.

Results suggest that the subsidy significantly changes the financing structure of corporate R&D. The subsidy has a statistically significant effect on firms at the bottom quartile and at the median of the self-financed R&D investment. The subsidy has a particularly strong effect for firms at the lowest quartile of self financed R&D. This effect remains positive and significant at the median. Again I find that the subsidy has no statistically significant effect for firms at higher quantiles. These results contrast with findings on externally financed R&D. As suggested by the model, firms at lower quantiles of total R&D investment do not access external funding before or after the subsidy. Consequently quantile estimates do not have a direct interpretation at lower quantiles. However the R&D subsidy seems to crowd out external finance for firms at higher quantiles of the R&D distribution. Finally estimating the relationship between the subsidy and external financial dependence I find a strong negative effect of the subsidy on firms at the median of the external financial dependence measure. The effect is statistically non-significant on firms at the bottom and the upper quartile of the distribution. Qualitatively similar results are found in the first robustness check where we estimate the equation only for firms with positive external financing of R&D (especially for the sign of the coefficient) although the precision of the estimates is affected in some cases. Censored quantile regression also shows a crowding out effect in terms of external finance but only up to the

[12] Since our sample consists of firms independent from business groups it is rather unlikely that they will receive transfers from other firms

75th quantile. A "'horizontal"' interpretation of these results is more difficult since firms at the lowest quartile of private R&D are not necessarily at the lowest quartile of the self-financed R&D and so on. Indeed a firm at the lowest quartile of private R&D investment could potentially finance all its R&D through external finance and therefore be ranked at higher quantiles. In order to check the validity of a horizontal interpretation I first have to prove that there is a strong correlation between the rankings of the units in the different distributions. I therefore use the Pearson correlation coefficient and Spearman's rank correlation coefficient tests. The use of Spearmans' rank correlation test allows to relax the linearity assumption on the Pearson correlation coefficient test. Results are reported in the appendix. Whatever relation is considered I always reject the null hypothesis that the distributions are independent. In all cases there is a significant positive association between the ranking in the different distributions. A horizontal interpretation of the results seems warranted[13].

Two key results are of special interest. First of all, subsequent to a subsidy, firms at lower quantiles are able to raise internally additional cash. Secondly, for firms at higher quantiles of the R&D distribution, the subsidy substitutes nearly entirely for external financing. I complement these results by investigating technological and financial characteristics of firms at the lowest quartile of the R&D distribution. Results are reported in Table 8[14]. The fact that subsequent to a subsidy firms at lower quantiles start injecting internal cash and possibly reducing on external finance could imply some form of financial market imperfection. This interpretation raises an important question. Why did the firm not finance the R&D projects from the beginning if they had enough internal finance to do so. One of the distinct characteristics of firms at lower quantiles is that they are not significantly smaller than firms at higher quantiles of the distribution, but they are significantly less R&D intensive. Consequently R&D activities compete with other investment opportunities for funding. The subsidy not only improves the NPV of the considered project, but also serves as a certification device within the firm. Consistent with this hypothesis is the evidence (although weak) that the subsidy represents a much larger part of

[13]As an additional robustness check I separately estimate the effect of the subsidy on total private R&D for firms with positive external financing. Results indicate again a simple additivity effect and even crowding out in some cases. Consequently the ordering is preserved for our quantile regressions since the complementary effect at lower R&D quantiles can be traced back to firms that report no external financing in their R&D activities.

[14]Considered outcomes of interest are the financing structure of R&D of the firms, as well as their technological characteristics.

$$Y = \alpha_0 + Quartile.1 \cdot \alpha_1 + X \cdot \alpha_2 + h(LBG) + Year + Industry \tag{14}$$

I estimate this equation via OLS, where Y is the considered outcome and Quartile.1 is a dummy indicating whether a firm belongs to the lowest quartile in the distribution of R&D investment. In specification (2) I include a vector of covariates X as well as year and industry fixed effects.

the R&D budget in firms at the lowest quantiles of the distribution. The second feature of interest is the effect of the subsidy at higher quantiles of the R&D distribution. The additive and even crowding out effect found at higher quantiles indicates that these firms are not credit constrained. Instead the public financing displaces the private financing of these firms nearly on a 1 to 1 basis. Since these firms are relatively more R&D intensive the certification effect of the subsidy is less important, i.e. innovation is an important part in the business strategy.

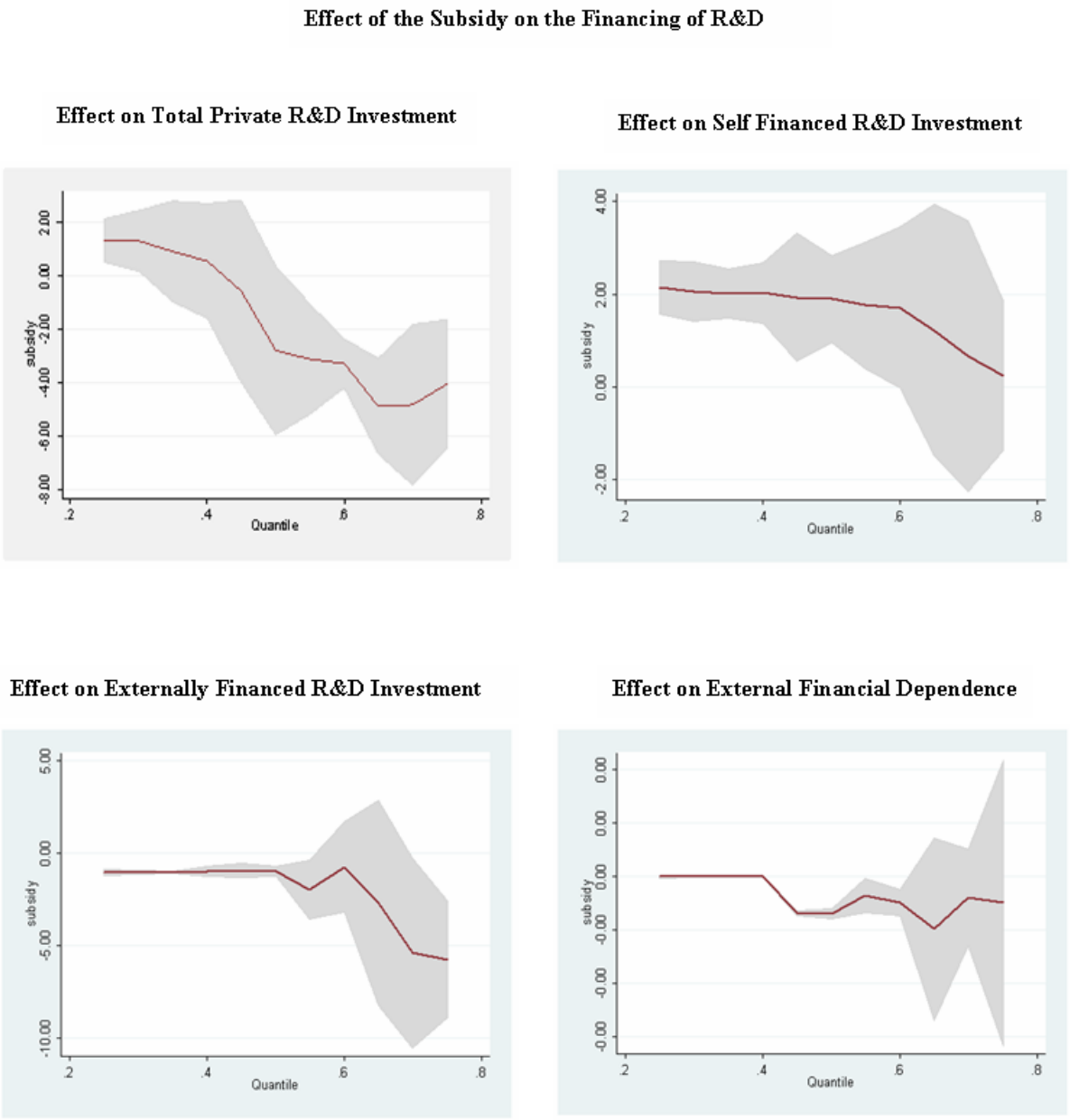


Figure 4: The Financing of R&D Investment

Table 7: The Financing Structure of Corporate R&D

		Quantile Regression Estimates				
	Bandwidth	Private R&D Investment	Internal Financing	Externally Financed	External Finance Ratio	Observations
.25 Quantile	Intermediate	1.17 (.32)	.04 (.42)	-.32 (.02)	-3.98x10(-10) (4.55x10(-09))	380
	Small	1.10 (.23)	1.27 (.59)	-1.06 (.008)	-2.26x10(-09) (2.21x10(-09))	276
	Very Small	1.33 (.4)	2.13 (.29)	-1.01 (.07)	-8.18x10(-08) (1.13x10(-07))	189
.5 Quantile	Intermediate	-1.55 (2.4)	1.69 (.65)	-.01 (.02)	-1.29x10(-07) (2.24x10(-07))	380
	Small	-.84 (1.5)	1.89 (.35)	-.12 (.09)	-2.75x10(-07) (2.65x10(-07))	276
	Very Small	-2.7 (1.6)	1.89 (.47)	-.96 (.14)	-6.9x10(-06) (4.97x10(-07))	189
.75 Quantile	Intermediate	2.55 (5.26)	-.4 (2.6)	-.42 (1.2)	-6.59x10(-07) (9.9x10(-06))	380
	Small	1.78 (4.93)	-.21 (1.8)	-1.5 (.73)	-1.04x10(-05) (4.96x10(-06))	276
	Very Small	-4 (1.21)	.23 (.81)	-5.73 (1.59)	-4.97x10(-06) (1.34x10(-05))	189

Each coefficient (and related standard error in parenthesis) is an estimate of the coefficient on the subsidy obtained from separate regressions of the form :

$$Y = \beta_0 + (\text{Subsidy})\beta_1 + X \cdot \beta_2 + \text{Year} + \text{Industry}$$

Where the dependent variable is indicated in each column and defined as :

Private R&D Investment (column 1) = Internal Financing of R&D (column 2) + External Financing of R&D (column 4)

$$\text{External Financing Dependence} = \frac{External Financing of R\&D}{Private R\&D Investment}$$

Subsidy is the actual level of treatment indicated in the corresponding row of the table;

X is a vector of covariates.

Estimates are obtained using quantile regression.

Year consists of Time Dummies. Industry consists of Industry Fixed Effects.

Table 8: Differences between Lower and Higher Quantiles

	Specification	
	(1)	(2)
Employment	-1864	-150
	(462)	(242)
Total Sales	-3748	-815
	(911)	(472)
Probability to Receive a Subsdiy	.03	.02
	(.14)	(.03)
Amount of Subsidy Received	-850	-1145
	(228)	(385)
R&D Intensity	-1.8	-2.2
	(.7)	(.95)
Fundamental Research	-.02	-.02
	(.01)	(.009)
Probability of Alternative Subsidy	-.07	-.066
	(.04)	(.04)
Covariates	No	Yes
Time Effects	No	Yes
Industry Effects	No	Yes
Number of Observations	560	560

Each coefficient (and related robust standard error in parenthesis) is an estimate of the coefficient on being at the lowest quartile of R&D investment
obtained from separate regressions of the form :
$Y = \alpha_0 + Quartile.1 \cdot \alpha 1 + X \cdot \alpha 2 + h(LBG) + Year + Industry$
Where Y is the pre-intervention outcome indicated in the corresponding row of the table;
Quartile1 is a dummy indicating whether a firm belongs to the group of firms at the lowest quartile of R&D investment.
Estimates are obtained using OLS.
Year stands for time dummies. Industry stands for Industry Fixed Effects.

Quantile Estimates: Self-Certification Hypothesis versus Scale Effects

Table 9: Threshold and Scale Effects

	Quantile Regression Estimates			
	.25 Quantile	.5 Quantile	.75 Quantile	Observations
Large Bandwidth	.04	.07	.11	560
$(0\% < X < 50\%)$	(.002)	(.009)	(.011)	
IntermediateBandwidth	.04	.06	.09	380
$(5\% < X < 45\%)$	(.004)	(.027)	(.043)	
Small Bandwidth	.05	.08	.23	276
$(10\% < X < 40\%)$	(.006)	(.02)	(.06)	
Very Small Bandwidth	.06	.25	.11	189
$(15\% < X < 35\%)$	(.02)	(.1)	(.02)	

Each coefficient (and related standard error in parenthesis) is an estimate of the interaction between the subsidy and employment obtained from separate regressions of the form :

Y=(Subsidy) β_1+(Subsidy)*(Non-R&D Employment) β_2 + Xβ_3 + Year + Industry

Where Y is the private net investment into R&D investment

Where (Subsidy*Non-R&D Employment) is the interaction term in the corresponding row of the table;

X is a vector of covariates from the fully specified OLS and IV-LATE model.

Estimates are obtained using quantile regression.

Year consists of Time Dummies. Industry consists of Industry Fixed Effects.

One major concern is that the analysis so far confused a potential self-certification effect with a simple scale effect in R&D activities. It seems plausible that R&D activities, because of their joint cost structure, exhibit to some degree scale effects. For instance, if thanks to the subsidy a firm is able to buy a high quality computer it will also positively affect the cost of other projects. Since such effects are more likely to occur at lower levels of R&D investment one would obtain similar predictions to the predictions under self-certification. One can try to distinguish between both hypothesis by using their differential prediction with respect to the size of a firm. Scale effects can indeed explain the quantile pattern observed between R&D investment and the effect of an R&D subsidy. However scale effects in R&D activities should then be independent from the size of the firm as measured by non-R&D employment. On the other hand, the self-certification hypothesis predicts that for a given quantile of R&D expenditures the certification effect of

the subsidy should be higher the larger the firm, i.e. the larger the non-R&D employment of the firm. Table 9 implements the outlined test by estimating the basic equation and interacting the amount of the subsidy with the non-R&D employment of the firm [15]. At all quantiles of the R&D distribution, the positive effect of the subsidy increases the larger the firm. The result is robust to bandwidth specification and statistically significant. I am therefore confident that results are not merely driven by scale effects in R&D activities.

[15] Alternatively I interacted the subsidy also with dummy indicators for different thresholds in terms of employment. Results are qualitatively the same.

8 Conclusion

The proposed model of self-certification starts from the insight that managers within organizations face imperfect information about the quality of their R&D activities. The severity of this informational problem decreases with the R&D intensity and investment of the firm. In firms where R&D is just of marginal importance to the business strategy, managers will have less incentives to be informed about R&D projects and their quality. Their imperfect information about the quality of their R&D activities leads them to invest less into innovation. Firms whose business model depends crucially on their ability to innovate will not face these internal doubts. If these firms want to do profitable business they have to innovate. Consequently managers of these "strong innovators" have an incentive to be perfectly informed about the quality and the activity of their R&D department. In this setting, the subsidy acts as a certification device within firms that are "marginal innovators", but has no informational content for the "strong innovators". In order to test the model, the study focuses on the effect of R&D support by the ANVAR agency for Small and Medium Sized firms. Estimation via Regression Discontinuity Design uses legal eligibility requirements to the ANVAR program to achieve identification in a quasi-experimental setting. Predictions obtained from this stylized certification model closely match empirical results. Overall, "marginal innovators" do not access external funding but "strong innovators" do. The explanation being that managers of "strong innovators" feel confident enough about the intrinsic quality of their R&D to justify accessing costly external funds. This implies that "strong innovators" invest the optimal amount of R&D, but "marginal innovators" don't. In this setting, the subsidy acts as a certification device within firms that are "marginal innovators", but has no informational content for the the other group. Consequently firms for which R&D is an important part of their business model use the subsidy as a risk sharing device, substituting public funds for costly external finance. For "marginal innovators" the subsidy acts as a self-certification device that triggers an increase in private R&D investment driven by internal funding. Complementary tests show that these results are neither driven by simple scale or threshold effects nor by external financial constraints. A goal of further research in this article will be to model in more detail the source of the imperfect information of managers in "marginal innovators".

References

[1] Aghion, P. and Howitt, P., 1998, Endogenous Growth Theory, MIT Press.

[2] Angrist, J.D. and Lavy, V., 1999, Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement, The Quarterly Journal of Economics, Vol.114(2), pp. 533 - 575.

[3] Arrow, K.J., 1962, Economic Welfare and the Allocation of Resources to Innovation, The Rate and Direction of Inventive Activity, Princeton University Press.

[4] Battistin, E., Rettore, E., 2003, Another Look At The Regression Discontinuity Design, Cemmap Working Paper, CWP01/03, pp. 1 - 19.

[5] Bennedsen, M., Wolfenzon, D., 200O, The Balance of Power in Closely Held Corporations, Journal of Financial Economics, Vol. 58, pp. 113 - 139.

[6] Buchinksy, M., 1994, "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression." Econometrica, Vol 62, pp. 405 - 458.

[7] Buchinksy, M., 1991, Methodological Issues of Quantile Regression, The Theory and Practice of Quantile Regression, Dissertation, Harvard University.

[8] Busom, I., 2000, An Empirical Evaluation of the Effects of R&D Subsidies, Economics of Innovation and New Technology, Vol. 9(2), pp. 111-148.

[9] Czarnitzki, D. and Fier A., 2004, Do Innovation Subsidies Crowd out Private Investment? Evidence from the German Service Sector, Applied Economics Quarterly, Vol. 48(1), pp. 1 - 25.

[10] David, P.A. and Hall B., 200O, Heart of Darkness: Modeling Public-Private Funding Interactions Inside the R&D Black Box, Research Policy, Vol. 29(9), pp. 1165 - 1183.

[11] David, P.A., Hall, B. and Toole, A.A., 200O, Is Public R&D a Complement or Substitute for Private R&D ? A Review of Econometric Evidence, Research Policy, Vol. 29(4-5), pp. 497 - 530.

[12] Duguet, E., 2004, Are R&D subsidies a substitute or a complement to privately funded R&D? Evidence from France using propensity score methods for non experimental data, Revue d'Economie Politique, Vol. 114(2), pp. 263 - 292.

[13] Garibaldi, P., Giavazzi, F., Ichino, A. and Rettore E., 2007, College Cost and Time to Complete a Degree: Evidence from Tuition Discontinuities, NBER Working Paper, N.12863.

[14] Guiso, L., 1998, High-Tech Firms and Credit Rationing, Journal of Economic Behaviour and Organization, Vol. 35, pp.39-59.

[15] Griliches, Z., 1998, R&D and Productivity: The Econometric Evidence, The University of Chicago Press.

[16] Hahn, J., Todd, P. and W. Van Der Klaaw (2001), Identification and Estimation of Treatment Effects with a Regression Discontinuity Design, Econometrica, Vol. 69(1), pp. 201 - 209.

[17] Helpman, E. and Grossman, G., 1991, Innovation and Growth in the Global Economy, MIT Press.

[18] Imbens, G. and Lemieux, T., 2007, Regression Discontinuity Designs : A Guide to Practice, NBER Working Paper, N.13039.

[19] Gonzalez, X., Jaumandreu, J. and Pazo, C., 2005, Barriers to Innovation and Subsidy Effectiveness, Rand Journal of Economics, Vol. 36(4), pp. 930-950.

[20] Klette, T. J. and Moen, J., 1998, R&D Investment Response to R&D Subsidies: a Theoretical Analysis and a Microeconometric Study, Paper presented at the NBER Summer Institute..

[21] Lach, S., 2000, Do R&D Subsidies Stimulate or Displace Private R&D? Evidence from Israel, Journal of Industrial Economics, Vol. 50(4), pp. 369 - 390.

[22] Lach, S. and Sauer, R. M., 2002, R&D, Subsidies and Productivity, Working Paper, pp. 1 - 47.

[23] Lichtenberg, F. R. 1984, The Relationship Between Federal Contract R&D and Company R&D, The American Economic Review, Vol. 74 (2), pp. 73 - 78.

[24] Lichtenberg, F. R. 1987, The Effect of Government Funding on Private Industrial Research and Development, The Journal of Industrial Economics, Vol. 36 (1), pp. 97 - 104.

[25] Lee, D.S., 2007, Randomized Experiments from Non-Random Selection in US House Elections, Journal of Econometrics, forthcoming.

[26] Klette, T.J., Moen, J. and Griliches Z., 2000, Do Subsidies to Commercial R&D Reduce Market Failures ? Microeconometric Evaluation Studies, Research Policy, Vol. 29, pp. 471 - 495.

[27] Wallsten, S., 2000, The Effects of Government-Industry R&D Programs on Private R&D : The Case of Small Business Innovation Research Program, Rand Journal of Economics, Vol. 31(1), pp. 82 - 100.

[28] Winston, C., 2006, Government Failure Versus Market Failure : Microeconomic Policy Research and Government Performance, Brookings Institutions Press.

Appendix: Data and Variables

Data The analysis combines 3 French firm-level databases: "Enquête Annuelle des Entreprises" managed the Ministry of Industry, "Enquête R&D" managed by the Ministry of Research and Education, and "Liaisons Financières" managed directly by INSEE. I construct a unified panel including information on firm characteristics (such as size, benefits..), R&D expenditures and financial links. The need for the information on financial ties between firms is due to the fact that eligibility is determined (at least partly) by the ownership status of a firm (subsidiary or independent). 1. "Enquête Annuelle d'Entreprise" is a yearly survey by the Ministry of Industry. It covers the whole manufacturing sector as well as the transport sector and the food processing industries (more than 24000 firms per year). Not only does it include fiscal and financial information such as revenues, benefits, cash flow and investments, but also information on employment. 2. "Enquête Annuelle sur les Moyens Consacrés à la R&D dans les Entreprises" is a yearly survey by the Ministry of Research. It covers a representative sample of all firms with more than 20 employees that undertake R&D (approximately 2000 firms per year). The database provides information on general characteristics of the firm, its innovative strategy (number of patents, type of patents), its internal R&D expenses (expenses, type of R&D investment, researchers) and external R&D expenses (subsidies and reimbursable aid by agency or ministry). 3. "Enquête Liasons Financieres" is a yearly survey by INSEE complemented with data from a legal database called DIANE. It covers all economic activities but restricts its attention to firms with either more than 500 employees, more than 30ME revenue, or a certain number of traded shares. LIFI gathers information about direct participations and shareholders. All databases are available for a long time horizon, however it appeared more practical to use only information from 1995 onwards (since data collection has been modified at several moments). Information on financial links was not available for one year (2003) and ownership status was assumed constant with respect to the year after for which data was available. Results are robust to the inclusion/exclusion of the given year.

Variable Definition

- Alternative Subsidies, total amount of public funds received from sources other than ANVAR;
- Applied Research, total R&D investment into applied research;
- Employment, amount of full-time employees;
- Experimental Research, total R&D investment into experimental research;

- Externally-Financed R&D Investment, amount of R&D investment financed by external partners of the firm (including banks);
- External-Financing Dependence, ratio Externally-Financed investment Net of Subsidy Private Net Investment;
- Fundamental Research, total R&D investment into fundamental research;
- High-Tech, whether a firms R&D intensity is above the median;
- LBG, percent of equity held by a business group of more than 2000 employees;
- Location, whether a firms' activity is located in a high-tech area (for instance "'Paris"', i.e. the region of Ile de France);
- Net of Subsidy R&D Investment, total R&D expenditures minus the subsidy received from ANVAR;
- Non-R&D Employment, total number of full-time employees minus the number of full-time researchers;
- Patent, total number of patents held by the firm;
- Process Innovation, whether a firm invests or not into process innovation;
- Product Innovation, whether a firm invests or not into product innovation;
- R&D Intensity, ratio of Net of Subsidy R&D Investment to Total Sales;
- Researchers, total number of full-time researchers;
- Sales, total sales by the firm;
- Self-Financed R&D Investment, amount of R&D investment financed by the firm;
- Subsidy, amount of subsidy from ANVAR received by a firm in a given year;
- Quartile1, whether a firm belongs to the lowest quartile in terms of Net of Subsidy R&D investment

Appendix: Robustness Checks

Table 10: Base Outcomes in Non-Treated Groups

	Mean Differences in R&D Investment (log)		
	Non-Treated Eligibles	Ineligibles	P-Value of Difference
Large Bandwidth	9.5	9.43	.7
Intermediate Bandwidth	9.33	9.47	.53
Small Bandwidth	9.2	9.34	.55
Very Small Bandwidth	9.49	9.52	.92

Two sample t test with equal variance
Null hypothesis of equality of means versus alternative hypothesis of inequality in means
Private R&D investment in logs

Table 11: Pearson Correlation Coefficient

	Pearson Correlation Coefficient		
	Net Private R&D	Self-Financed R&D	Externally Financed R&D
Net Private R&D	1		
Self-Financed R&D	0.91 (.00)	1	
Externally Financed R&D	0.65 (.00)	.31 (.00)	1

Pearson correlation coefficient with significance levels in parenthesis
Null hypothesis of independence between the two distributions

Table 12: Spearman Correlation Coefficient

	Pearson Correlation Coefficient		
	Net Private R&D	Self-Financed R&D	Externally Financed R&D
Net Private R&D	1		
Self-Financed R&D	0.68 (.00)	1	
Externally Financed R&D	0.56 (.00)	-.01 (.71)	1

Spearman correlation coefficient with significance levels in parenthesis
Null hypothesis of independence between the two distributions

Table 13: Robustness Checks for External Financing

	Subsample of Positive External Finance		Censored Quantile Regression	
	Externally Financed	External Finance Ratio	Externally Financed	External Finance Ratio
.4 Quantile	1.02 (2.46)	-3.98x10(-4) 5.5x10(-4)	-	-
.6 Quantile	-.37 (3.7)	-5.6x10(-4) 1.61x10(-4)	-2.12 (.21)	-3.79x10(-4) 1.54x10(-4)
.8 Quantile	.76 (8.3)	-.913x10(-4) 13.1x10(-4)	1.7 (2.2)	2.6x10(-3) 2.3x10(-4)

Each coefficient (and related standard error in parenthesis) is an estimate of the coefficient on the subsidy obtained from separate regressions of the form :

$$Y=g(LBG)+ \beta(Subsidy) + (X) + Year + Industry$$

Where the dependent variable is indicated in each column and defined as :

External Financing of R&D (columns 2 and 4)

External Financing Dependence $= \frac{External Financing of R\&D}{Private R\&D Investment}$ column (3 and 5)

Subsidy is the actual level of treatment indicated in the corresponding row of the table;

X is a vector of covariates from the fully specified OLS and IV-LATE model.

Estimates in comumns 1 and 2 are obtained using quantile regression on the subsample of firms with external financing of R&D

Estimates in columns 4 and 5 are obtained using censored quantile regression methods using Buchinsky (1991, 1994)

If estimates in columns 4 and 5 did not converge at the indicated quantile the next highest .05 quantile was used. Otherwise estimate not indicated

Year consists of Time Dummies. Industry consists of Industry Fixed Effects.

Market Structure and the Diffusion of Electronic Banking*

Jason Allen
Bank of Canada

Robert Clark
HEC Montréal

Jean-François Houde
University of Wisconsin-Madison

February 12, 2008

Abstract

This paper studies the role that market structure plays in affecting the diffusion of electronic banking. Electronic banking represents a process innovation since it reduces the cost of performing many types of transactions for banks. However, electronic banking (and electronic commerce more generally) is particular since the full benefits for firms from adoption only accrue once consumers begin to perform a significant share of their transactions online. Since it is costly for consumers to switch to the new technology (they must learn how to use it) banks may try to encourage consumers to go online by affecting the relative quality of the online and offline options. Their ability to do so is a function of market structure since in more competitive markets, reducing the relative attractiveness of the offline option involves the risk of losing customers (or potential customers) to competitors, whereas, this is less of a concern for a more dominant bank. Based on the Beggs and Klemperer (1992) model of price competition, we develop a model of branch-service quality choice with switching costs meant to characterize the trade-off banks face when rationalizing their network between technology penetration and business stealing. The model is solved numerically and we show that the incentive to lower branch-service quality and drive consumers into electronic banking is greater in more concentrated markets and for more dominant banks. We find support for the predictions of the model using a panel of household survey data on electronic payment usage as well as branch location data, which we use to construct a measure of branch quality (namely branch density).

PRELIMINARY. PLEASE DO NOT CITE

JEL classification: D14, D4, G21, L1

*Correspondence to Jean-François Houde: University of Wisconsin-Madison, Madison, Wisconsin; Phone: (608) 262-3805; Email: houdejf@ssc.wisc.edu, Robert Clark: HEC Montréal, CIRANO and CIRPÉE, Montreal, Quebec; Phone: (514) 340-7034; Email: robert.clark@hec.ca, and Jason Allen: Bank of Canada, Ottawa, Ontario; Phone (613) 782-8712; Email: jallen@bankofcanada.ca. Robert Clark thanks the FQRSC and HEC Montréal for funding. We thank Micromedia Proquest as well as Peter Kascor at Credit Union Central of Canada for information on credit unions. We thank the Big 6 Canadian banks for giving us access to IT executives who patiently answered many of our questions. We thank Gautam Gowrisankaram and Randy Wright for comments and suggestions as well as seminar participants at the Bank of Canada, the Competition Bureau, and HEC Montréal. The views in this paper do not necessarily reflect those of the Bank of Canada. All errors are our own.

1 Introduction

This paper studies the diffusion of electronic banking in the retail banking industry. By electronic banking we mean the set of tools allowing consumers to perform most of their day-to-day banking transactions remotely through the internet. Electronic banking represents a cost-reducing technology since for many types of day-to-day transactions it is cheaper for banks if consumers perform them online.[1] We are interested in the role that market structure plays in affecting this diffusion. Understanding the effect of market structure on diffusion in this industry is important since in recent decades there has been considerable consolidation in retail banking markets throughout the world, and evidence suggests this trend will continue into the future.

The relationship between market concentration and the diffusion of a new process innovation (a technology that reduces the cost of production) has been studied extensively. The focus of this literature is on the trade-off that firms face between the incentive to delay adoption, since the cost of adoption is expected to fall over time, and the incentive to adopt early in order to prevent or delay the adoption by competitors in the case of strategic rivalry.[2] Competition may speed up diffusion since it encourages a preemptive technology adoption motive. In the literature it has been common to assume that once firms adopt the new technology, any increase in returns is immediately realized. There are instances, however, where the realization of the full benefits from the introduction of a new technology depends on the extent to which consumers use it rather than the old technology. In the day-to-day banking market, despite the fact that banks have adopted electronic payment mechanisms, the realization of the full benefits from these mechanisms depends on the decisions of consumers to perform transactions electronically rather than at traditional bricks-and-mortar branches. This is true in general for innovations in electronic commerce.[3]

The fact that diffusion is consumer driven potentially implies a different role for market structure in affecting firm incentives and the resulting diffusion of new technologies such as e-commerce that,

[1] For instance, using internal data from 20 of the top U.S. banks, Boston Consulting Group (2003) concludes that banks could double profits if customers switched from offline to online bill payment. Also, DeYoung, Lang, and Nolle (2007) report a positive correlation between community bank profitability and early adoption of an operational website.

[2] See Reinganum (1981a), Reinganum (1981b), and Fudenberg and Tirole (1985) for theoretical analyzes of the effect of market concentration on the speed of adoption. Kamien and Schwartz (1982) survey the early empirical work looking at this relationship. See also early work by Levin, Levin, and Meisel (1987), Hannan and McDowell (1984), and Karshenas and Stoneman (1993). More recently this question has been studied by Hamilton and McManus (2005), Schmidt-Dengler (2006), Gowrisankaran and Stavins (2004) (for technologies featuring network externalities), and Seim and Viard (2006).

[3] Another example is self-serve kiosks at airports and self-check-out kiosks at grocery stores. Airlines/grocery stores may invest in the installation of electronic kiosks, but the benefits from adoption for firms, are only realized once consumers start checking in/checking out electronically.

to our knowledge, has not been studied. There has, however, been some work examining the effect that e-commerce has on market structure. For instance, Emre, Hortaçsu, and Syverson (2006) look at the effect of the introduction of e-commerce on market reorganization in a number of industries. They find that in the auto dealer and book store industries small stores exited local markets where the use of e-commerce channels grew fastest. But the underlying assumption in their analysis is that the diffusion of e-commerce is an exogenous process. This may not be an appropriate assumption in markets where firms operate both online and offline channels. In such markets firms may have an incentive to affect the relative attractiveness of online versus offline transactions in order to encourage consumers to adopt the less costly technology. Evidence suggests that offline price and the local availability of offline outlets can affect the use of electronic commerce by consumers (see Goolsbee (2000), Prince (2006), and Forman, Ghose, Goldfarb (2006)). Therefore banks may try to encourage consumers to switch to the new technology by adjusting the relative prices of online and offline banking and/or by reorganizing their retail networks (apparently this was the approach employed by banks in Scandinavia to encourage consumers to switch to online banking (The Economist, June 14th 2007)).

The ability of firms to make these adjustments depends on the level of competition in the local market. There is evidence that competition plays a role in affecting banks' reorganization decisions. For instance, Cohen and Mazzeo (2005) analyze the effect of market structure on branching decisions and find that branch networks are larger in more competitive markets. Therefore, in more competitive markets, reducing the attractiveness of traditional retail stores by closing branches involves the risk of losing customers (or potential customers) to competitors, whereas, this is less of a concern for a more dominant bank. In the case of e-banking, instead of encouraging a pre-emptive technology adoption motive, increased competition generates a business stealing effect, slowing the penetration of the cost-reducing technology.[4]

We develop a dynamic model of branch-quality competition that characterizes the tradeoff banks face between (i) making branch banking relatively less attractive to encourage consumers to switch to electronic banking – we refer to this as the *technology penetration* incentive –, and (ii) maintaining quality for fear of losing consumers to rivals – we refer to this as the *business stealing* incentive. The model generates testable predictions about the effect of competition on usage/adoption of electronic banking. We find that competition tends to increase the quality of branch networks offered by banks

[4]The relative concentration of banking markets in Scandinavian countries has been put forth as an explanation for the high rates of adoption of other types of electronic payment technologies (Milne 2005).

and therefore decrease the usage rate of electronic transactions. This prediction is in contrast to that found in the literature that has examined the relationship between market concentration and the diffusion of a new process innovation. As mentioned above, in contrast with our hypothesis, the traditional view is that adoption is typically faster in more competitive markets since competition encourages a preemptive technology adoption motive.

Our empirical analysis focuses on the Canadian retail banking industry. Over the past decade, the largest Canadian banks have profoundly changed their way of offering retail banking services. The Canadian industry features a small number of large banks that traditionally provided an extensive network of branches for their clients. However, between 1998 and 2006 the top eight Canadian banks on average reduced the number of retail branches they operated by 21 per cent.[5] In December 1997, The Royal Bank of Canada became the first Canadian bank to offer some banking services online and soon after the major Canadian banks all had online operations. Canadians have quickly become among the world's heaviest users of electronic payments. The number of transactions performed electronically increased from 47 million to over 300 million from 2000 to 2006 (Canadian Bankers Association), while the share of consumers who did at least some online banking increased from 3 per cent in 1998 to 43 per cent in 2006.

In order to study the substitution between online and offline banking channels and the role that branch quality and market structure play in affecting this substitution we combine two unique data sets. The first is the Canadian Financial Monitor (CFM) database compiled by Ipsos-Reid Canada. This data set contains information on the usage of different banking channels in the period immediately following the introduction of online banking in Canada (1999-2006), along with detailed information on the demographic characteristics of respondents. To measure the quality of the branch network we use location data from the "Financial Services Canada" directory produced by Micromedia Proquest. The directory provides information on branch locations in all local markets for all of the years in our sample as well as years prior to the introduction of electronic banking.[6] With this information we construct measures of branch density (number of branches per capita) to reflect the quality of the offline option since there is convincing evidence that consumers care strongly about the extent of a bank's network of branches and automated bank machines (ABM's) (See Kiser (2004), Bernhardt and Massoud (2004), and Grzelonska (2005). In the case of Canada, a recent study found that 56 per cent of respondents chose a bank because if its convenient branch

[5]To be precise, it is the top eight banks other than TD Bank Financial Group which we exclude from this measure since it closed many branches as a result of the 2000 merger with Canada Trust.

[6]For the most part, we will define a local market to be a census division, of which there are 288 in Canada.

and ABM network (Deutsche Bank (2005))).[7]

Our empirical work supports the prediction that banks can rationalize their networks in order to encourage adoption and that it is easier to do so in less competitive markets and for more dominant banks. We first show that initial market structure affects the change in the number of branches per capita in the market. In more concentrated markets and in markets with more dominant banks there are more branch closures. Having shown this, we confirm that this translates into an effect on e-banking by establishing that a significant relationship exists between branch closures (or changes in the number of branches per capita) and e-banking usage. We study this relationship first at the market level and then we provide further evidence by performing a household-level analysis in which we consider the effect of changes in branch density in a household's local neighbourhood on their usage and adoption of e-banking. We show that branch closures cause increased usage and adoption.

We conclude that initial market structure and branch network reorganization have an effect, therefore, on e-banking usage. Our results do not suggest that the mechanism described in Emre, Hortaçsu, and Syverson (2006), whereby firms reorganize their retail network in response to the diffusion of e-commerce, does not exist. Rather, we provide evidence of an additional incentive to reorganize one's retail network. In markets such as banking, where firms offer both an online and offline channel, closures can encourage adoption.

The paper proceeds as follows. Section 2 provides a condense overview of the Canadian banking industry. Section 3 presents a model of quality competition with switching costs. Section 4 presents our empirical analysis. Section 5 concludes.

2 The Canadian banking market

The Canadian retail banking industry features a small number of very large federally regulated national institutions that dominate most local markets.[8] The industry is best described as stable (Bordo 1995) with almost no exit, and little entry, at least on the retail side of banking.[9] The major

[7]We could also look at operating hours or number of tellers. However, number of branches affects wait times and travel distances while these other quality measures affect only wait times.

Relative prices could also have an effect in some banking markets, but not at the local market level since the Canadian retail banking industry features a small number of very large national institutions that dominate most local markets and so although for consumers day-to-day banking is done locally, banking fees are common across regions.

[8]These banks are Royal Bank Financial Group, Bank of Montreal, Canadian Imperial Bank of Commerce, TD Bank Financial Group, and Bank of Nova Scotia.

[9]There has been a large inflow of foreign banks into the Canadian market but mostly on the corporate side of banking. A few foreign banks have made inroads in the retail market, most notably ING Canada, a virtual bank.

banks provide similar products and services and are not dis-similar in terms of standard measures of productivity and efficiency (Allen and Engert 2007). There has been one substantial merger during our sample. In 2000 TD Bank and Canada Trust merged to become TD Bank Financial Group (something we control for in our empirical analysis).

The industry is characterized by several key facts: (i) 85 per cent of banking assets are held by the five largest banks; (ii) deposits at these institutions are growing; (iii) at least one of these banks operates in 98 per cent of the census divisions, and at least two in 81 per cent;[10] (iv) and branches are being closed. The remainder of the Canadian banking industry is characterized by a large number of small banks, both foreign and domestically owned, as well as provincially regulated credit unions. Some credit unions have a strong presence in a particular set of local markets and are therefore important to include in our analysis. Examples include Caisse Desjardins (Quebec), ATB Financial (Alberta), and Vancity (British Columbia).

As previously mentioned, the Canadian banking industry is relatively concentrated. A Figure of the Herfindahl-Hirschman indices (HHI) averaged across census divisions and smoothed using a kernel estimator is presented in the appendix for 1998.[11] There is a large mass slightly over 2000 as well as a substantial mass beyond that, indicating a high degree of concentration in many markets.

Over the past decade, the largest Canadian banks have profoundly changed their way of offering retail banking services. Between 1998 and 2006 the top eight Canadian banks have on average reduced the number of retail branches they operate by 21 per cent, despite a 37 per cent increase in deposits.[12,13] In this period Canadians have quickly become among the world's heaviest users of electronic payments. The number of transactions performed electronically increased from 47 million to more than 300 million from 2000 to 2006, while the share of consumers who did at least some online banking increased from 3 per cent in 1997 to 43 per cent in 2006. We also know through a number of different surveys that the majority of Canadian consumers are satisfied with the provision of new banking technologies (83 per cent of Canadians reported in 2004 of being either satisfied or

[10]There are 288 census divisions in Canada. A census division corresponds roughly to a municipality or a county. The largest is Toronto with more than 2.5 million individuals and the smallest is Stikine with 1100 individuals. Census divisions are used by Statistics Canada to conduct Canada's census every five years.

[11]We define the HHI of a market j as the sum of market shares squared, where the market share of bank i, for example, is the fraction of branches owned by bank i in market j. In many U.S. studies of banking, deposits at the branch level are usually taken us the measure of market share. Given data restrictions we can only tabulate total deposits for each bank at the provincial level. As one would expect, however, the number of branches controlled by a bank in a province and the value of deposits by that bank are highly correlated, with a correlation coefficient of 0.9.

[12]In contrast, in the period from 1982-1997 the top six Canadian banks closed only 2.3% of their branches.

[13]Branch closures are frequently in the Canadian news. Changes to the Bank Act in 2002 established the Financial Consumer Agency of Canada,that publicizes branch closures and provides consumers with information on what do to if their branch intends to close.

very satisfied), and the reason they bank online is convenience (in 2004 78 per cent of Canadians said they adopted because online banking was more convenient).[14]

3 Model

In the literature studying the adoption of process innovations firms must decide when to incur the cost of adopting a new technology. The focus has been on the trade-off that firms face between the incentive to delay adoption, since the adoption cost is expected to fall over time, and the incentive to adopt early in order to prevent or slow the adoption by competitors in the case of strategic rivalry. Adoption should therefore be faster in more competitive markets.

In the context of markets where the benefits from a new technology only accrue once consumers have switched to it, the primary 'adoption cost' that firms must incur is the cost of encouraging consumers to switch. In other words, banks devote resources to making it more attractive for consumers to engage in e-banking (so we can think of these resources as spending on promotion or on enhancing the quality of the website).

Rather than making the new technology more attractive, an alternative mechanism via which banks can encourage penetration of the new technology is to make the old technology less attractive by reducing the quality of branching service. The aim of this section is to contrast the impact of these two mechanisms on the diffusion of ebanking. To do so, we develop a model of bank competition with switching costs based on Beggs and Klemperer (1992) in which consumers must decide where to bank and what fraction of their day-to-day transactions to perform online, and banks can influence these decisions in one of two ways: (i) by spending an amount Q_o to make the online option more attractive for consumers (we will refer to this as the Online-Quality mechanism), or (ii) by reducing the quality of branching services Q_b (we will refer to this as the Branch-Quality mechanism).[15]

In each of infinitely many discrete time periods two banks non-cooperatively and simultaneously choose either the quality of online service (Online-Quality mechanism) or the quality of branch service (Branch-Quality mechanism) to provide to their customers in an effort to maximize their total expected future discounted profits. In each period a cohort of new consumers enters the market to join a group of old consumers. Old consumers have already bought banking services in earlier

[14] Canadian Bankers Association, "Technology and Banking: A Survey of Canadian Attitudes 2004."

[15] Of course, in reality banks might make use of both of these mechanisms simultaneously. We do not permit them to do so since the goal of this section is to contrast the outcomes that arise when banks use the two mechanisms.

periods and are assumed to never switch away from the bank they patronized in previous periods.[16] Competition, therefore is in order to attract new consumers.

When banks employ the Online-Quality mechanism they have incentive to spend on Q_o for two reasons. First, doing so increases the utility of consumers (by making online banking more attractive) and therefore ultimately increases a bank's market share (in other words, spending on online quality has a positive influence on the *business stealing* effect). Second, doing so lowers its costs since a greater proportion of transactions will be done via the less expensive technology (spending on online quality also has a positive influence on the *technology penetration* effect).

In contrast, when they employ the Branch-Quality mechanism they face a tradeoff when reducing the quality of branching services between *technology penetration* and *business stealing*. By lowering quality they attract a lower share of new consumers. However, at the same time, since consumers must decide on the fraction of their transactions to perform online versus at a branch, lower quality branching service encourages more use of the online channel on the part of consumers, reducing costs.

In order to analyze the effect that competition has on these incentives to devote resources to improving the quality of the online option or to lowering the quality of the offline option we consider the effect of adusting the cost of switching. If switching away from a bank is more costly, competition is reduced since consumers are more captive. We are interested in determining the effect of changing the cost of switching on steady-state online or offline quality levels and resulting usage rates. The model is developed as follows and is solved numerically.

3.1 Branch-Quality mechanism

The problem of old consumers affiliated with a bank of branch quality Q_b is to choose the proportion of transactions to be performed online, denoted by μ, by trading off the relative cost of e-banking over teller-based transactions. This problem is static, and with a probability $(1 - \rho_j)$ a customer of

[16] Dube et al. (2006) set up a model in which all consumers are able to switch. We think that the fact that there is no switching is not restrictive in our case since, as we show in Section 4 below, there are very few switches observed in the data.

bank j will be allowed to switch away. The utility maximization problem is the following:

$$u(Q_b) = \max_{\mu} \quad \gamma + (1-\mu)(Q_b - p_b) + \mu(-p_e) - \frac{\lambda}{2}\mu^2 \tag{1}$$

$$\Leftrightarrow \quad \mu(Q_b) = \frac{p_b - p_e - Q_b}{\lambda}, \tag{2}$$

where $p_b - p_e > 0$ is the price differential between transactions performed at a branch (teller) and transactions performed electronically, and λ represents a technological-familiarity parameter (the bigger is λ the less familiar with or less able to access technology are consumers). It is useful to write the indirect utility function as a function solely of $\mu(Q_b)$, by replacing $Q_b(\mu) = p_b - p_e - \lambda\mu$ such that:

$$u(\mu) = \delta - p_e - \lambda\mu + \frac{\lambda}{2}\mu^2. \tag{3}$$

The problem of new consumers is first to decide which bank to patronize, and then the proportion of transactions performed online. New consumers are assumed to be uniformly distributed along the unit line, and a consumer located at i must incur a "transportation" cost $t|i-j|$ to choose a bank located at point j. Consumers have two banks from which to choose. Bank 0 is located at 0, while bank 1 is located at 1. Demand for each bank is determined by an indifferent type, $z(\mu_0, \mu_1)$:

$$z(\mu_0, \mu_1) = \frac{\lambda(\mu_1 - \mu_0) + \frac{\lambda}{2}(\mu_0^2 - \mu_1^2)}{2t} + \frac{1}{2} \tag{4}$$

The firms' problem is a dynamic game in quality (or equivalently in the proportion of online-transactions, μ_j). Assuming that firms base their strategies only on current payoff relevant state variables (i.e. Markov strategies), the Bellman equation of bank 0 is given by:

$$V_0(x|Q_1^b) = \max_{\mu_0} \quad (\frac{F(x|\mu_0,\mu_1)}{\rho_0})\big[(1-\mu_0)(p_b - c_b) + \mu_0(p_e - c_e)\big] - \frac{C}{2}Q_b(\mu_0)^2 + \delta V_0(F(x|\mu_0,\mu_1)|\mu_1), \tag{5}$$

where $p_e - c_e > p_b - c_b$ (i.e. the markups on electronic transactions is higher than on teller transactions) and where $F(x|\mu_0,\mu_1) = ((1-\rho_0)x + (1-\rho_1)(1-x))z(\mu_0,\mu_1) + \rho_0 x$ represents bank 0's stock of old consumers next period if its current stock is x (a fraction ρ_0 of its current stock do not exit (switch) and it captures a fraction $z(\mu_0,\mu_1)$ of the exiters (switchers) from both banks ($(1-\rho_0)x$ of its own switchers and $(1-\rho_1)(1-x)$ from bank 1)). The first term in (5) represents bank 0's current revenue from the two channels since current period sales are given by $\frac{F(x|\mu_0,\mu_1)}{\rho_0}$ (we divide

by ρ_0 to condition on the survival rate at bank 0). The problem of bank 1 is defined symmetrically, replacing x by $1-x$ and z by $(1-z)$.

Differentiating (5) with respect to μ_0 we obtain the first order condition for bank 0's equilibrium level of usage:

$$0 = (\frac{1}{\rho_0}\frac{\partial F(x|\mu_0,\mu_1)}{\partial \mu_0})\big[(1-\mu_0)(p_b-c_b)+\mu_0(p_e-c_e)\big]$$
$$+(\frac{F(x|\mu_0,\mu_1)}{\rho_0})(p_e-c_e-(p_b-c_b))-C\frac{\partial Q_b(\mu_0)}{\partial \mu_0}+\delta\frac{\partial V_0(F(x|\mu_0,\mu_1))}{\partial F(x|\mu_0,\mu_1)}\frac{\partial F(x|\mu_0,\mu_1)}{\partial \mu_0}$$

where $\frac{\partial F(x|\mu_0,\mu_1)}{\partial \mu_0} = ((1-\rho_0)x+(1-\rho_1)(1-x))\frac{\partial z(\mu_0,\mu_1)}{\partial \mu_0}$. From the first order condition we can see the tradeoff banks face when reducing the quality of branching services between *technology penetration* and *business stealing*. The first term represents the business stealing effect and is negative since $z(\mu_0,\mu_1)$ is decreasing in μ_0 (increasing quality causes usage to decrease and market share to increase). The second term represents the technology penetration effect and is positive since when μ_0 increases more transactions are performed using the more profitable channel. Note also that since greater usage is associated with lower quality, the third term is positive.

3.2 Online-Quality mechanism

Rather than lower Branch-Quality, banks can adjust Online-Quality by choosing how much to spend on Q_o. The consumer problem then becomes:

$$u(E) = \max_{\mu} \quad \gamma+(1-\mu)(-p_b)+\mu(Q_o-p_e)-\frac{\lambda}{2}\mu^2 \tag{6}$$
$$\Leftrightarrow \quad \mu(Q_o) = \frac{P_b-P_e+Q_o}{\lambda}. \tag{7}$$

Writing the indirect utility function solely as a function of $\mu(Q_o)$ (by replacing $Q_o(\mu) = -P_b+P_e+\lambda\mu$) we can solve for the indifferent new consumer

$$z(\mu_0,\mu_1) = \frac{\lambda(\mu_0^2-\mu_1^2)}{4t}+\frac{1}{2}.$$

Using this, we can write bank 0's Bellman equation as follows:

$$V_0(x|\mu_1) = \max_{\mu_0} \quad (\frac{F(x|\mu_0,\mu_1)}{\rho_0})\big[(1-\mu_0)(p_b-c_b)+\mu_0(p_e-c_e)\big] - \frac{C}{2}Q_o(\mu_0)^2 + \delta V_0(F(x|\mu_0,\mu_1)|\mu_1). \tag{8}$$

Differentiating (8) with respect to μ_0 we obtain the first order condition for bank 0's equilibrium level of usage:

$$\begin{aligned} 0 \quad = \quad & (\frac{1}{\rho_0}\frac{\partial F(x|\mu_0,\mu_1)}{\partial \mu_0})\big[(1-\mu_0)(p_b-c_b)+\mu_0(p_e-c_e)\big] \\ & +(\frac{F(x|\mu_0,\mu_1)}{\rho_0})(p_e-c_e-(p_b-c_b)) - C\frac{\partial Q_o(\mu_0)}{\partial \mu_0} + \delta\frac{\partial V_0(F(x|\mu_0,\mu_1))}{\partial F(x|\mu_0,\mu_1)}\frac{\partial F(x|\mu_0,\mu_1)}{\partial \mu_0}. \end{aligned}$$

In contrast with the first order condition given above when banks use the Branch-Quality mechanism, from the first order condition for the Online-Quality mechanism we observe that the *technology penetration* and *business stealing* effects operate in the same direction. When banks use the Online-Quality mechanism $z(\mu_0,\mu_1)$ is increasing in μ_0 (increasing online quality causes usage to increase and market share to increase). The technology penetration effect is also positive since when μ_0 increases more transactions are performed using the more profitable channel. Note here that that since greater usage is associated with higher online quality, the third term is negative.

3.3 Model Results

We solve the model numerically. To do so we follow Beggs and Klemperer (1992) and assume that the value function of the banks takes a known parametric form. Since the function $z(\mu_0,\mu_1)$ is quadratic in the decision variable of firms (instead of linear as in Beggs and Klemperer (1992)), we conjecture that the value function will be a cubic function of the state variable x. The solution of the problem then involves finding values for the parameters of the value functions that satisfy the Bellman and Nash conditions.

The numerical values for the parameters used to compute the solution are given in Table 1. Our qualitative results hold as long as the profit from an e-banking transaction (π_e) is greater than for a branch transaction (π_b) and that the consumer price of an e-transaction is less than that same transaction performed at a branch.

The results of the numerical exercise are summarized in Figure 1, which shows steady-state

Table 1: Numerical values for the model parameters

Technological familiarity:	λ	$[1.5, 3]$
Bank fixed cost:	C	2
Switching cost:	ρ_j	$\{0.5, 0.8\}$
Branch price:	p_b	1.25
E-banking price:	p_e	0.5
Branch transaction profit:	π_b	0.25
E-banking transaction profit:	π_e	0.5
Utils from banking:	γ	1
Unit transportation cost:	t	1/4
Discount factor:	δ	0.8

usage rates when banks employ the two mechanisms for different values of λ (i.e. the technological familiarity parameter) and ρ_j (i.e. the switching cost). The top two figures characterize what happens when banks employ the Branch-Quality mechanism, the bottom two characterizes behaviour for the Online-Quality mechanism. In each figure, the solid line represents the usage in the situation where switching costs are symmetric across banks ($\rho_0 = \rho_1 = \rho$), while the dotted and the dashed lines are usage of the firms with high and low switching costs respectively. The first thing to note is that, for both mechanisms and regardless of the cost of switching, as λ falls, usage increases. This is not surprising as we would expect online usage to increase as the cost of performing online transactions falls.

First, we investigate the effect of decreasing the level of competition in the market. We consider the situation where the cost of switching is symmetric across banks and examine what happens as ρ increases. In this case, using the Branch-Quality mechanism we observe that as ρ increases (moving from the left panel to the right panel), usage increases. This is because in less competitive markets branch quality is lower and this generates higher usage. The opposite is true when banks use the Online-Quality mechanism. As ρ increases, we see that usage decreases. In less competitive markets online quality is lower and usage is lower. What is going on is that as ρ increases, the business-stealing effect becomes less important relative to the technology-penetration effect since consumers are more captive. With the Branch-Quality mechanism the only thing preventing banks from lowering quality is the fear of losing customers to rivals via the business-stealing effect. And this effect becomes less important as ρ increases. In contrast, with the Online-Quality mechanism, banks have a double incentive to increase quality since the two effects work in the same direction. As ρ increases, the incentive to increase quality to steal customers from rivals is diminished and so

Online-Quality is lower, and therefore usage is as well.

Second, we study the effect of increasing the dominance of one of the banks. We consider the situation with asymmetric switching costs. For the Branch-Quality mechanism we find that the bank with the higher switching cost generates higher usage. Since its switching cost is higher, it worries less about losing customers to its rival and so can afford to lower quality resulting in higher usage. Again, the opposite is true for the Online-Quality mechanism. The bank with the lower switching cost has higher usage, implying that weaker firms choose higher online quality. Again, as ρ_j increases, the business-stealing effect becomes less important relative to the technology-penetration effect.

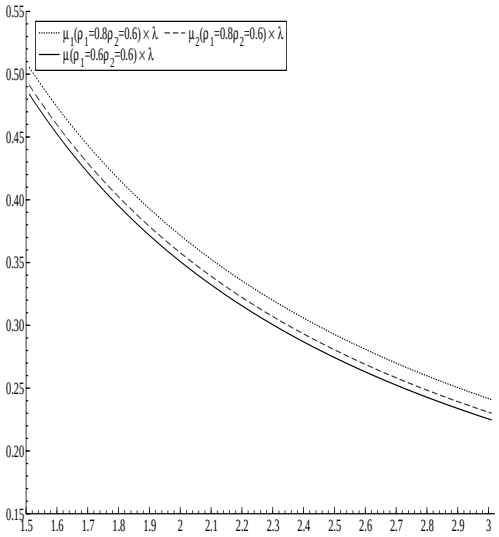
We summarize our results in the following proposition

Proposition 1. *The following comparative static results obtain:*

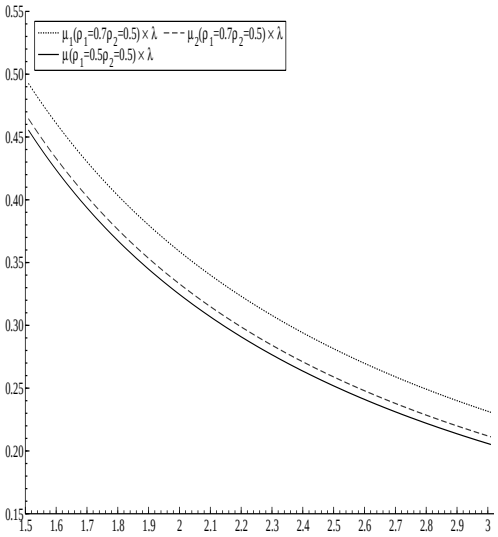
1. *Suppose the cost of switching is symmetric across banks ($\rho_0 = \rho_1 = \rho$), then*
   - *if using the Branch-Quality mechanism, in less competitive markets (higher ρ) quality is lower and usage is higher.*
   - *if using the Online-Quality mechanism, in less competitive markets (higher ρ) quality is lower and usage is lower.*
2. *Suppose the cost of switching is asymmetric across banks, then*
   - *if using the Branch-Quality mechanism, a bank that faces less competition (higher ρ_j) will have lower quality and higher usage.*
   - *if using the Online-Quality mechanism, a bank that faces less competition (higher ρ_j) will have lower quality and lower usage.*

Note that since the relationship between usage and market structure predicted for the two mechanisms is different, the relationship between branch closures and market structure will be as well. That is, the Online-Quality mechanism implies less e-banking usage in less competitive markets and so if closures are the result of increased adoption and usage of e-banking (and not the cause thereof), there should actually be fewer closures in less competitive markets when banks employ the Online-Quality mechanism.

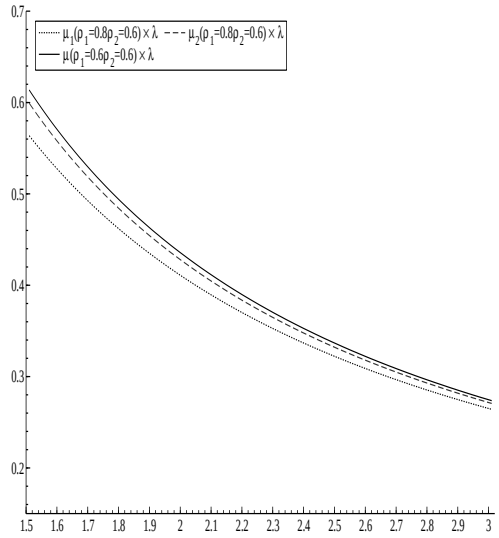
Figure 1: Steady state usage rates



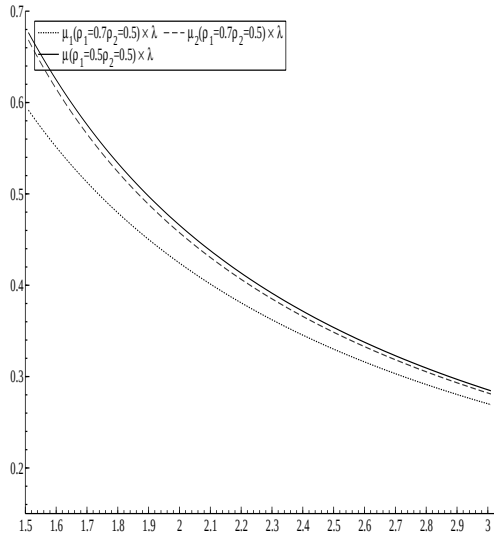
(a) Branch quality model with high switching cost



(b) Branch quality model with low switching cost



(c) Online quality model with high switching cost



(d) Online quality model with low switching cost

4 Empirical Analysis

In this section we present empirical evidence that suggests that banks employ the Branch-Quality mechanism (rather than the Online-Quality mechanism). The model predicts that if using the Branch-Quality mechanism, banks that operate in less competitive markets or that are dominant will lower branch service quality in order to encourage consumers to use the online channel. To test this prediction we combine two unique data sets. The first contains information on the usage of different banking channels, along with detailed information on the demographic characteristics of respondents. The second contains the location information of all branches in our sample period and is used to construct a measure of branch density with which we proxy for branch-service quality. We describe these data sets below before turning to our empirical results.

4.1 Data

4.1.1 Canadian Banking Habits

We use detailed consumer-level data characterizing household decisions to adopt electronic payment technologies as well as banking relationships and detailed demographic characteristics. This is done by combining Census information with household financial data obtained from the "Canadian Financial Monitor" (CFM) survey results compiled by Ipsos-Reid.

We use the complete survey results – 1999 to 2006. On average there are approximately 12,000 Canadians surveyed per year (staggered evenly by quarter), with a non-trivial number of individuals staying in the survey for more than 1 year and up to 8 years.[17] The geographical distribution of households in the survey is similar to the total population across all census divisions (CDs), where each census division is labeled a market.

Survey responses provide us with a substantial amount of information regarding household characteristics. In our analysis we focus on those characteristics which are most likely to be correlated with bank channel choice.[18] Helpful in this choice are results previously documented by Stavins (2001) who showed, using the limited data available in the 1998 U.S. Survey of Consumer Finances that internet bill payments were more likely to be conducted by younger households, those with

[17] There are a total of 76204 people in the sample. Of these, we observe 24 113 just once, 15 600 twice, 11 238 three times, 8 676 four times, 6 645 five times, 4 764 six times, 3 360 seven times, and 1 808 eight times.

[18] The survey provides a wealth of information on household assets and liabilities which could be used as controls beyond our current analysis.

high income and home ownership, those with better education and those who hold white collar jobs. Summary statistics are presented in Table 2. Summary statistics are conditioned on the respondent's sex –which, the majority of time, is female (approximately 76 per cent).

Table 2: Summary of Household Characteristics: 1999-2006

CHARACTERISTIC	Mean	Median	Std. Dev
Respondent: age†	46.7	46	14.9
Respondent: education	15.3	14	2.5
Maximum: age	51.9	51	15.1
Maximum: education	15.7	16	2.5
Household: income($)	61,568	57,500	35,581
Household: size	2.5	2	1.3
Duration: primary bank*	11.1	12	4.9
Transaction cost‡($)	5.67	2.5	7.4

Note:†The age variable refers to the age of the respondent in 1999. Respondents under the age of 18 in 1999 are dropped. This represents only 0.02 per cent of the sample. *Duration is right-censored at 20 years therefore we report the average duration for those reporting less than 20 years, which represents close to 50 per cent of the sample. ‡Transaction costs are almost entirely unreported in the panel prior to 2004. The reported figures are for households surveyed after 2003 and defined as service charges paid in the last month.

From Table 2 we notice immediately that the average duration of a banking relationship is relatively long, the median is 20+ years. Given the high proportion of households that have a relationship with their bank exceeding 20 years we speculate that switching costs are relatively high. Focusing on those households that are seen repeatedly in the sample, we find that 3.1 per cent of these switch out of their main financial institution to either an institution previously recorded as secondary or a new institution.[19]

Table 3 documents the number of households using each of the possible banking options. With respect to usage rates, we find that the majority of households continue to visit a teller at least once a month, but this number has fallen over time as more households adopt e-banking. The usage rates for phone banking in this paper is the major reason we do not include phone banking and e-banking as a single alternative to branch banking. Phone-banking is a mature delivery channel and there are not many new adopters in our sample. The number of transactions, conditional on making a transaction has not moved very much over time. Also, although usage rates are lowest for e-banking, households that use this technology make a large number of transactions. The share of ABM, teller, and phone transactions have all fallen over time. Noticeably, therefore, the share

[19]More conservatively, we find that 1.25 per cent of households record switching to a new bank.

of PC-transactions has increased substantially over the sample period, from 4.2 per cent to 19.5 per cent of total transactions. Table 4 breaks down the e-banking activities of Canadians into four main categories. The majority of e-banking is for day-to-day purposes, typically bill payment and transfers. Online banking is therefore a substitute for teller-banking. The second most popular use of banking websites is to gather information. This includes gathering information on mortgages, investments, and credit cards. Most Canadians do not perform credit or investment activities online.

Table 3: Summary of Banking Channel Usage

TYPE	1999	2000	2001	2002	2003	2004	2005	2006
				Adoption rates				
Respondent: PC at work	52.7	58.1	67.7	71.0	72.0	72.5	75.0	75.7
Maximum: PC at work	58.2	62.4	71.1	74.1	75.3	75.7	77.9	78.3
Teller	82.8	80.7	78.0	77.1	77.0	76.4	71.8	75.4
ABM	72.0	71.6	72.3	73.0	71.8	71.2	70.9	69.8
Phone	30.3	31.7	32.3	31.6	30.6	30.6	30.3	29.2
PC	13.4	17.3	25.8	32.5	34.7	36.8	41.3	42.8
				Share of Total Transactions				
Teller	27.8	28.1	26.6	25.7	25.7	26.4	24.8	26.1
ABM	57.5	55.5	54.1	53.0	51.0	48.8	48.7	46.5
Phone	10.5	10.3	10.0	9.4	9.3	9.5	8.2	8.2
PC	4.2	5.9	9.2	11.9	14.0	15.3	18.3	19.2

Note: Rates and shares are reported in percentage points.

Table 4: Summary of e-banking Activities

Activity	2001	2002	2003	2004	2005	2006
Share day-to-day	66.2	69.7	72.9	75.2	76.5	77.1
Share information gathering	24.8	22.2	18.9	16.2	14.6	14.7
Share credit	3.8	3.9	4.0	4.0	4.5	4.2
Share investment	5.2	4.2	4.1	4.5	4.4	3.9

Note: Usage rates and shares are reported in percentage points.

4.1.2 Branch Density

Our measure of bank quality is the density of its branch network. This seems like a realistic approximation given the empirical evidence provided in Kiser (2004), Bernhardt and Massoud (2004), and Grzelonska (2005). Branch location information on all financial institutions in Canada has been scanned and transferred to electronic files from the "Financial Services Canada" directory produced

by Micromedia Proquest. The directory is cross-listed with branch information provided by the Canadian Payments Association, branch-closing dates reported by the Financial Consumer Agency of Canada, branch closing and opening information provided in the annual reports of Canada's largest banks (a process that started in 2002 via the Accountability Act), and location data provided directly by some of the banks. In what follows we provide a description of the data.[20]

At the market level we want to examine the impact of density variables on bank-channel adoption. Summary statistics are reported in Table 5. The average number of branches in a market is 4 per square kilometer and 5.7 per 100 000 people. The average change in branches per capita (*dbranchcap*) is -21 per cent. The average change in branches per square kilometer (*dbranchdens*) is -17 per cent. Rationalization of branches (most precisely measured as *dbranchdens*) is consistently high for the different group sizes, although highest for the largest banks. We include these variables in the regression analysis reported in section 4.2.[21]

Table 5: Summary of Bank Statistics: 1998-2006

	Total		Large		Medium		Small	
VARIABLE	Mean	SD	Mean	SD	Mean	SD	Mean	SD
branchdens	4.0	14.9	11.6	24.6	0.69	0.33	0.11	0.08
dbranchdens	-0.17	0.43	-0.22	0.43	-0.14	0.44	-0.14	0.41
branchcap	5.71	6.65	12.5	9.1	4.27	0.79	1.73	0.83
dbranchcap	-0.21	0.43	-0.18	0.50	-0.19	0.35	-0.24	0.45

Note: We present the mean and standard deviation (SD) for four groupings: total as well as large (biggest third), medium, and small census divisions. Branch density is in banks per square kilometer and Branches per capita is in branches per 100 000 people.

In our analysis we must control for the acquisition of Canada Trust Financial Services by Toronto-Dominion Bank, now called TD Canada Trust or TD Bank Financial Group. TD completed its $8 billion acquisition on February 1st, 2000.[22] With the acquisition TD acquired approximately 600 branches. We can assume that many of these branches were closed to save costs. Similarly we can

[20] At the time of this paper we do not have access to all of the banks ABM network, limiting the analysis to branch location choice. A substantial fraction of brand-name ABM machines (as opposed to white-label machines), however, are located in branches. Also, according to our CFM survey, more than 60 per cent of ABM transactions are at the branch, a number that has been slowly increasing since 2001. This is likely because of the change in composition of ABMs from largely brand labels to white-labels.

[21] Given the historical development of the Canadian banking industry data is collected by the regulator and supervisor at the national (sometimes provincial) level. This is why, unlike in the United States, we do not have branch-specific data on deposits, number of employees, and branch-specific investment in capital.

[22] Throughout the 1990's, due to a change in regulation, the large Canadian banks acquired most of the Trust companies. These Trust companies were both smaller than Canada Trust and were acquired prior to our sample start date.

assume some of the TD branches were closed in favour of keeping open a more efficient Canada Trust branch. Fortunately TD Bank has provided us with a list of closures, including dates, for these type of branches. We therefore control for closures by TD Bank that are likely to be merger-related.

4.1.3 General Market Characteristics

In addition to household survey data and branch location information we include in our analysis general characteristics of the cross-section of local markets. To characterize our markers we use 2001 and 2006 census data on population, age, and employment. Summary statistics on key variables are reported in Table 6. We use this information to control for local market activities which might affect a bank reorganization decisions.

4.2 Analysis

The theoretical model presented above predicts that if using the Branch-Quality mechanism, banks that operate in less competitive markets or that are dominant will lower branch-service quality in order to encourage consumers to use the online channel. We test this prediction by proxying for branch service quality with the number of branches per capita in the market, and by studying the relationship between market structure, branch-service quality, and diffusion of e-banking. As mentioned above, we define a market as being a census division of which there are 288.

We start by studying the effect of initial market structure on branch-network rationalization to confirm that banks operating in less competitive markets and more dominant banks have a greater incentive to lower quality. Having shown this, we confirm that this translates into an effect on e-banking by establishing that a significant relationship exists between branch closures (or changes in the number of branches per capita) and e-banking usage. We study this relationship first at the market level and then provide further evidence by performing a household-level analysis in which we consider the effect of changes in branch density in a household's local neighbourhood on their usage of e-banking.

4.2.1 Effect of initial market structure on changes in branch-service quality

At the market level, defined at the census division level (288 markets), we study factors influencing the change in branch service quality, (proxied for by branches per capita). In order to control for the

year 2000 merger between TD Bank and Canada Trust we attribute all TD Canada Trust closures to the merger. In effect, we assume that TD Bank's decision to close branches was never in order to encourage its consumers to adopt online banking.

Table 7 presents regression results for the change in the number of branches per capita in market m ($branchcap_m$) over the sample period on market structure variables:

$$\log(\frac{branchcap_{m06}}{branchcap_{m98}}) = \theta HH98_m + \lambda nbcomp98_m + Z_m\gamma + \epsilon_m, \tag{9}$$

where $HH98_m$ is the initial (1998) level of concentration of all the banks in the market, $nbcomp98_m$ is the initial number of competitors in the market, and Z_m is a vector or market variables that includes the average age of individuals living in the market, their average income, and the average employment level.

From column (1) we see that the initial market structure variables, $HH98_m$ and $nbcomp98_m$, are both negative and significant which implies that when the market is initially more concentrated, more branches are closed. Controlling for the initial number of banks, an increase in the initial Herfindahl index implies that the market is less competitive. Controlling for the initial Herfindahl index, an increase in the number of competitors makes the market less competitive in the sense that it implies the existence of at least one more dominant firm. These results provide empirical evidence in support of the Branch-Quality mechanism. The number of branches per capita is smaller in more concentrated markets.

Columns (2) through (5) of Table 7 include controls for changes in PC banking or PC/Phone banking (Home banking) usage and/or adoption levels during the sample period. We can see that the market structure result does not change.[23] We discuss the relationship between e-banking and branch closures in further detail in Section 4.2.3 below.

In Table 8 we present regression results for the change in the number of bank j's branches per capita in market m ($branchcap_{jm}$) over the sample period on market structure variables:

$$\log(\frac{branchcap_{jm06}}{branchcap_{jm98}}) = \alpha share98_{jm} + \theta HH98_{jm} + \lambda nbcomp98_m + Z_m\gamma + \epsilon_{jm}, \tag{10}$$

[23] We instrument for changes in e-banking usage or adoption with change in web access since this variable is highly correlated with e-banking usage and adoption but should not affect closures independently. Note that there are only 84 observations in these regressions since we can only calculate a reliable measure of e-banking usage rates for 84 of the census divisions in 1998.

where $share98_{jm}$ is bank j's own initial share of market m and $HH98_{jm}$ is the initial level of concentration amongst j's rivals in the market. Our results provide further support for the second prediction of the model, that if using the Branch-Quality mechanism, more dominant banks have more incentive to lower branch-service quality. We find that a larger initial market share is associated with more branch closures. This result is consistent regardless of specification. We consider three different specifications to capture the effect of rival attractiveness/competitiveness. From column (3) we see that the more rivals bank j has initially ($nbcomp98_m$) the more it closes over the sample period. This is because, given j's market share, the more rivals j has, the fewer branches each has, thus making them less attractive. Similarly, in column (2) we report the effect of the Herfindahl index of bank j's rivals in 1998 ($HH98_{jm}$). The more concentrated are j's rivals, the fewer branches j closes. In column (1) we control for the Herfindahl index and the number of rivals simultaneously. When doing so, the coefficient on the Herfindahl index is no longer significant while the coefficient on ($nbcomp98_m$) is still negative and significant. The interpretation of this coefficient is different than when ($nbcomp98_m$) enters on its own. Controlling for the initial Herfindahl index, an increase in the number of rivals implies the existence of at least one more dominant rival for bank j. One might therefore expect this coefficient to be positive and for bank j to close fewer branches, but it may be that rivals are less attractive to consumers on average if one is quite large and others are small, or that the more dominant rival is more attractive and branch closures are strategic complements (if j faces a more attractive rival and its rival closes more branches, j can close more branches also).

4.2.2 Effect of changes in branch-service quality on e-banking usage and adoption

We know from our first set of regressions that initial market structure affects closures. Having shown this, we confirm that this translates into an effect on e-banking. We do so by establishing that a significant relationship exists between branch closures (or changes in the number of branches per capita) and e-banking usage. We study this relationship first at the market level and then provide further support by performing a household-level analysis in which we consider the effect of changes in branch density in a household's local neighbourhood on their usage of e-banking.

In Table 9 we report results for the following regression

$$\log(\frac{ebanking_{m06}}{ebanking_{m98}}) = \beta \log(\frac{branchcap_{jm06}}{branchcap_{jm98}}) + Z_m\gamma + \epsilon_m, \tag{11}$$

where $ebanking_{mt}$ is either PC or PC and Phone (Home) banking usage or adoption in market m

in period t. We test the effect of the change in the number of branches per capita in the market on the change in e-banking usage and adoption rates. We find that initial market structure affects the change in Home banking usage and adoption but does not have a significant effect on the change in PC banking. These results suggest that the closures that occur in less competitive markets are driving consumers into both PC and Phone banking. The results are qualitatively similar if we instrument for closures using the initial market structure variables ($HH98$ and $nbcomp98$).

To confirm that e-banking usage depends on closures we look deeper into the data to determine whether at the household level, branch density influences the decision to use e-banking. As mentioned above our measure of branch quality is branch density. Branch density is made household-specific by counting the number of branches of a particular household's bank in a circle with a 1 kilometer radius around the centroid of that household's postal code (nbh). The mean number of own-bank branches per neighborhood of this type is 0.44 with a variance of 0.82. The minimum is zero and the maximum 21.

Parameter estimates from the following Tobit regressions are estimated for the share of PC and Home banking. Parameter estimates are reported in Table 10.

$$share^*_{it} = \max(0, \theta nbh_{ijt} + X_{ijt}\beta + Z_{jt}\gamma + \epsilon_{ijt}), \quad share^*_{it} = \begin{cases} \text{Share-pc} \\ \text{Share-Home} \end{cases}, \tag{12}$$

where $share^*_{it}$ is household i's usage of either PC or Home banking in period t, the X_{it} are household control variables, and the Z_{it} market control variables. We find that PC and Home usage are both negatively correlated with the bank-branch density variable.[24,25] The result is qualitatively the same as we extend the size of a household's neighborhood.

Another advantage to the household level analysis is that it allows us to address the simultaneity bias that may exist since not only may branch closures lead to adoption and usage of e-banking by consumers, but adoption and usage of e-banking by consumers (or the anticipation thereof) may lead

[24] Our results regarding the impact of the various demographic variables are consistent with those reported in Stavins (2001).

[25] We have come across one other paper that looks at the effect of distance to branch on adoption of electronic banking, Khan (2004). Our results differ from Khan (2004) along a number of dimensions. Most importantly, Khan finds that distance does not matter for adoption. However, we have a much larger and richer data set. For example, we know the location of each of the household's bank's branches in their neighborhood which allows us to construct a measure of quality that captures the density of the branch network. Khan only uses the reported "distance to main branch" as the hypothesized explanatory variable. We also find that younger Canadians are more likely to adopt online banking than older Canadians. Khan finds that older Americans are the more likely adopters. This result is hard to rationalize given what is known about the adoption of new technologies more generally - younger individuals are more willing to try new technologies.

to branch closures. To address this problem we restrict attention to the sub-sample of consumers whose main financial institution was TD Bank or Canada Trust. Most TD or CT branch closures during our sample period were the result of the merger of these two institutions and were not the result of e-banking. If, following the merger, branches were located within two kilometers of each other, generally one was closed down. PC and Home usage are both still negatively correlated with branch density in the restricted sample.

We also test whether usage changes as a function of branch closures. We estimate the following regression:

$$\Delta share_{ijt} = \theta \Delta nbh_{ijt} + X_{ijt}\beta + Z_{jt}\gamma + \epsilon_{ijt},\ share_{ijt} = \begin{cases} \text{Share-pc} \\ \text{Share-Home} \end{cases} \tag{13}$$

and present results in Table 11. We find that a change in the number of branches inside of a household's local neighbourhood is correlated with a change in PC usage. Column (3) includes only TD and CT customers and the results are unchanged.

4.2.3 Effect of e-banking diffusion on market structure

It is important to note that our results are consistent with Emre, Hortaçsu, and Syverson (2006). From Table 7 we can see that in markets where e-banking is adopted there are more branch closures. Emre, Hortaçsu, and Syverson (2006) assume, however, that the diffusion of e-commerce is an exogenous process. We show that this may not be the case, at least for e-banking and for markets where firms operate both online and offline channels. Our results suggest that firms can affect e-banking usage and adoption by closing branches.

Our results do not suggest that the mechanism described in Emre, Hortaçsu, and Syverson (2006), whereby firms reorganize their retail network in response to the diffusion of e-commerce, does not exist. Rather, we provide evidence of an additional incentive to reorganize one's retail network. In markets such as banking, where firms offer both an online and offline channel, closures can encourage adoption. If closures occur simply in response to the diffusion of e-banking, then we should not observe more closures in markets that are less competitive initially. If the diffusion is exogenous, then there is no reason to believe that it will be a function of initial market structure. And if it is not a function of initial market structure, then the resulting closure pattern should not be either. Similarly, if diffusion is a result of improvements in the online option (Online-Quality

mechanism), then it should be faster in more competitive markets, and therefore the closures that might result from the fact that consumers are less in need of branches would also occur faster in more competitive markets.

5 Conclusion

In this paper we have studied the relationship between market structure and the diffusion of electronic banking. In the day-to-day banking market, despite the fact that banks have adopted electronic payment mechanisms, the realization of the full benefits from its introduction depends on the decisions of consumers to perform electronic transactions. This is true in general for innovations in electronic commerce. This paper sheds light on how banks can affect the relative attractiveness of their offline and online channels to encourage consumer adoption of innovations in e-banking. In particular, we show that banks can encourage online adoption by rationalizing their branch network.

A further contribution of this paper is that we show that the ability to rationalize branches depends on market structure in a non-standard way. We show that there are more closures in the most concentrated markets and it is the larger banks that close the most branches. The reason banks do this is to encourage adoption (technology penetration incentive) and they are able to do this in less competitive environments because the business stealing incentive is less binding. These results, therefore, provide empirical evidence to support the Branch-Quality model of competition presented in the paper.

In future work we extend the analysis to take into account for a number of features currently missing. For example, we currently fix the cost of adoption of e-banking for all consumers. A more realistic approach is to allow this cost to vary according to household characteristics and to the diffusion of internet technologies more generally. This would allow us to measure the welfare costs associated with the introduction of e-banking and with bank closures across low and high adoption cost households.

References

Allen, J. and W. Engert (2007). Efficiency and competition in canadian banking. *Bank of Canada Review*, Summer 2007.

Beggs, A. and P. Klemperer (1992). Multi-period competition with switching costs. *Econometrica*, 651–666.

Bernhardt, D. and N. Massoud (2004). Endogenous ATM location and pricing. mimeo.

Bordo, M. (1995). Regulation and bank stability: Canada and the united states, 1870–1980. World Bank Working Paper No. 1532.

Boston Consulting Group (2003). Online bill payment: A path to doubling profits.

Cohen, A. and M. Mazzeo (2005). Investment strategies and market structure: An empirical analysis of bank branching decisions. Mimeo, Federal Reserve Board.

Deutsche Bank (2005). E-banking snapshot.

DeYoung, R., W. Lang, and D. Nolle (2007). How the internet affects output and performance at community banks. *Journal of Banking & Finance*, 1033–1066.

Emre, O., A. Hortaçsu, and C. Syverson (2006). E-commerce and the market structure of retail industries. NBER working paper 2005-24.

Fudenberg, D. and J. Tirole (1985). Preemption and rent equalization in the adoption of new technology. *The Review of Economic Studies 52*(3), 383–401.

Gowrisankaran, G. and J. Stavins (2004). Network externalities and technology adoption: Lessons from electronic payments. *Rand Journal of Economics*, 678–693.

Grzelonska, P. (2005). Benefits from branch networks: Theory and evidence from the summary of deposits data. Manuscript.

Hamilton, B. and B. McManus (2005). Technology adoption and market structure: Evidence from infertility treatment markets. mimeo, Olin School of Business.

Hannan, T. and J. McDowell (1984). The determinants of technology adoption: The case of the banking firm. *Rand Journal of Economics*, 328–335.

Kamien, M. and N. Schwartz (1982). *Market Structure and Innovation*. Cambridge: Cambridge University Press.

Karshenas, M. and P. Stoneman (1993). Rank, stock and order effects in the diffusion of new process technologies: An empirical model of adoption duration. *Rand Journal of Economics*, 503–528.

Khan, B. (2004). Consumer adoption of online banking: Does distance matter? University of California at Berkeley working paper E04-338.

Kiser, E. (2004). Predicting household switch behavior and switching costs at depository institutions. *Review of Industrial Organization*, 349–365.

Levin, S., S. Levin, and J. Meisel (1987). A dynamic analysis of the adoption of new technology: the case of optical scanners. *Review of Economic Studies*, 12–17.

Milne, A. (2005). What's in it for us? network effects and bank payment innovation. Bank of Finland discussion paper 16-2005.

Reinganum, J. (1981a). Market structure and the diffusion of new technology. *Bell Journal of Economics*, 618–624.

Reinganum, J. (1981b). On the diffusion of new technology: A game theoretic approach. *Review of Economic Studies*, 395–405.

Schmidt-Dengler, P. (2006). The timing of new technology adoption: The case of mri. working paper.

Seim, K. and B. Viard (2006). The effect of market structure on cellular technology adoption and pricing. working paper.

Stavins, J. (2001). Effect of consumer characteristics on the use of payment instruments. *New England Economic Review*, 19–31.

Table 6: Summary of a Few Market (Census Division) Characteristics: 2001, 2006

	2001	2006
Census:		
Population		
mean	106079	111639
median	39196	39765
sd.	253527	267142
Income		
mean individual	25461	
median individual	25089	
sd individual	4233	
mean household	55776	
median household	54786	
sd household	9921	
Age		
mean share under 20	21.4%	20.1%
mean share 20-24	6.3%	6.1%
mean share 25-34	12.4%	11.6%
mean share 35-49	26.2%	24.1%
mean share 50-64	18.8%	22.1%
Education		
share high school degree or less	42.4%	
share with a degree	25.6%	
share with university degree	20.6%	
Occupation		
share management	8.3%	
share business/finance/administration	13.8%	
share sales/services	22.9%	

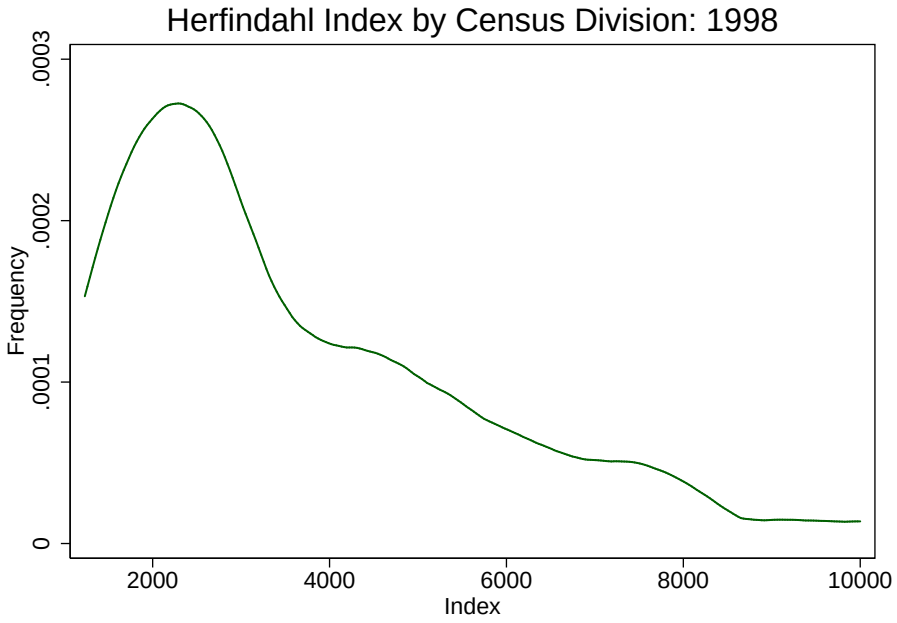

Herfindahl Index by Census Division: 1998
Frequency
0
.0001
.0002
.0003
2000
4000
6000
8000
10000
Index

Table 7: The Change in the Number of Branches per Capita

COEFFICIENT	LABELS	(1) D.Branch	(2) D.Branch (IV)	(3) D.Branch (IV)	(4) D.Branch (IV)	(5) D.Branch (IV)
HH98	Market HH index (1998)	-1.499***	-1.563**	-1.375***	-2.533***	-1.358***
		(0.16)	(0.63)	(0.51)	(0.42)	(0.40)
nbcomp98	Number of banks (1998)	-0.0725***	-0.0573***	-0.0512***	-0.0677***	-0.0502***
		(0.012)	(0.015)	(0.016)	(0.013)	(0.015)
dpop	Pop. change (2006/1998)	-0.803***	-0.355	-0.268	-0.371	-0.492*
		(0.28)	(0.29)	(0.29)	(0.27)	(0.28)
age	Age (2001)	-0.0431***	0.0134	0.0149	0.0152	0.0186
		(0.015)	(0.011)	(0.013)	(0.011)	(0.014)
avgincome	Avgerage income (2001)	-2.005**	-2.178**	-1.932***	-1.807**	-1.563**
		(0.86)	(0.96)	(0.73)	(0.73)	(0.59)
avgemp	Employment (2001)	0.00283	0.0109**	0.0111**	0.00885*	0.0118**
		(0.0027)	(0.0049)	(0.0051)	(0.0048)	(0.0054)
dshare_pc	Change in pc usage (2006/1998)		-0.114			
			(0.071)			
dhomeadopt	(mean) dhomeadopt					-0.368**
						(0.17)
dpcadopt	(mean) dpcadopt				-0.153**	
					(0.068)	
dshare_home	Change in home usage (2006/1998)			-0.301*		
				(0.16)		
Observations		276	84	84	82	84
R^2		0.39	0.36	0.43	0.43	0.45

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 8: The Change in the Number of Bank j's Branches per Capita

COEFFICIENT	LABELS	(1) dbranchcap	(2) dbranchcap	(3) dbranchcap	(4) dbranchcap
share98	Branch share in 1998	-0.556***	-0.348***	-0.541***	-0.308***
		(0.095)	(0.093)	(0.095)	(0.090)
HHi98	Competitors' HH in 1998	-0.121	0.161***		
		(0.078)	(0.059)		
nbcomp98	Nb. competitors in 1998	-0.0530***		-0.0444***	
		(0.0088)		(0.0067)	
dpop	Pop. change (2006/1998)	-0.721***	-0.763***	-0.721***	-0.776***
		(0.15)	(0.15)	(0.15)	(0.15)
age	Age (2001)	-0.00888	-0.00822	-0.00863	-0.00852
		(0.0055)	(0.0057)	(0.0055)	(0.0056)
avgincome	Avg. income (2001)	0.307	-0.323	0.335	-0.600*
		(0.40)	(0.39)	(0.40)	(0.36)
avgemp	Employment (2001)	0.0000924	0.000694	0.000274	0.000514
		(0.0015)	(0.0015)	(0.0015)	(0.0015)
Observations		1116	1116	1116	1116
R^2		0.44	0.42	0.44	0.42

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 9: Change in E-Banking Usage/Adoption as a Function of the Change in the Number of Branches Per Capita

COEFFICIENT	LABELS	(1) D.pc	(2) D.home	(3) D.pcadopt	(4) D.homeadopt
dweb	Change in web access (2006/1998)	1.609***	0.554***	1.509***	0.424***
		(0.49)	(0.16)	(0.23)	(0.14)
dtotbranchcap	Change in total branch per capita (2006/1998)	-0.319	-0.443***	0.0409	-0.486***
		(0.40)	(0.14)	(0.25)	(0.17)
dpop	Pop. change (2006/1998)	0.211	0.237	-0.243	-0.466
		(1.20)	(0.44)	(0.76)	(0.45)
age	Age (2001)	-0.0272	-0.000127	0.00354	0.0133
		(0.058)	(0.024)	(0.036)	(0.021)
avgincome	Avgerage income (2001)	-7.831***	-2.601***	-3.072**	-1.289
		(2.64)	(0.89)	(1.50)	(0.78)
avgemp	Employment (2001)	0.0314*	0.0149	0.0210**	0.0151
		(0.019)	(0.010)	(0.010)	(0.0096)
Observations		84	84	82	84
R^2		0.35	0.33	0.43	0.29

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 10: Household Level E-banking Usage Rates - Tobit

COEFFICIENT	LABELS	(1) PC usage	(2) Home usage	(3) PC usage	(4) Home usage
nbh2	Nb. branchs in 1km nbh.	-0.0437**	-0.0968***	-0.0873*	-0.121**
		(0.017)	(0.020)	(0.045)	(0.049)
web	Web access	0.391***	0.259***	0.399***	0.264***
		(0.013)	(0.013)	(0.030)	(0.029)
age	Age (in 1999)	-0.00186***	-0.00602***	-0.00204**	-0.00565***
		(0.00044)	(0.00047)	(0.00092)	(0.0010)
avgschool	Average HH schooling	0.00639***	0.00854***	0.00605	0.0100*
		(0.0023)	(0.0025)	(0.0050)	(0.0056)
Constant		-0.709***	0.116*	-0.577***	0.0728
		(0.068)	(0.065)	(0.11)	(0.11)
Observations		8067	8067	1573	1573
R^2		.	.	.	.

Tobit estimates. Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Columns 3 and 4 are estimated on the sub-sample TD/CT consumers.
All specifications also include occupation dummies and year/bank fixed effects

Table 11: Household Level Change in E-banking Usage Rates

COEFFICIENT	LABELS	(1) D.Home usage	(2) D.PC usage	(3) D.PC usage
Dnbh2	Change in 1 Km nbh.	-0.0364	-0.0550**	-0.0796*
		(0.038)	(0.025)	(0.047)
web	Web access	0.0315***	0.0267***	0.0372***
		(0.0079)	(0.0039)	(0.0080)
age	Age (in 1999)	0.000312	0.0000450	0.0000521
		(0.00031)	(0.00016)	(0.00030)
avgschool	Average HH schooling	0.00107	0.0000487	0.0000696
		(0.0016)	(0.00089)	(0.0021)
Constant		-0.117**	-0.0377	0.00265
		(0.050)	(0.025)	(0.035)
Observations		3626	3626	727
R^2		0.01	0.02	0.04

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All specifications also include occupation dummies, year/bank fixed effects

Column 3 includes only CT and TD consumers

AN EMPIRICAL MODEL OF MULTI-VENUE TRADING COMPETITION POST-MiFID

Ricardo Ribeiro*

March 2008

Abstract

The Market in Financial Instruments Directive (MiFID) aims to increase competition and to foster client protection in the European financial market. Among other provisions, it abolishes the concentration rule and challenges the market power of existing trading venues. The directive introduces venue competition in order to achieve better execution and ultimately lower trading costs. In this paper I address the question of whether fostering competition between alternative trading venues alone may or not be able to impact actual competition in the market. I consider two reasons for why it may not: cash trading exhibits direct network effects and the typical trading and post-trading bundling in the EU. I then propose an empirical framework to evaluate the actual degree of competition between trading venues. This empirical approach constitutes, for the best of my knowledge, one of the first attempts to structurally model financial trading, which is instrumental for measuring empirically the impact of network effects and of the bundle of trading and post-trading services as barriers to competition. This evaluation is provided in the companion paper, Ribeiro (2008).

JEL Classification: C13, G10, L11, L84

Keywords: Market Dominance, Network Effects, Bundling, Barriers to Competition, Demand Estimation, Financial Trading

*STICERD, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Tel: +44 20 7955 6690. Fax: +44 20 7955 6951. Email: r.c.ribeiro@lse.ac.uk. http://personal.lse.ac.uk/ribeiror/. Thanks are due to Peter Davis and John Van Reenen for their guidance and continuous help as well as to David Lawton, Carlos Santos and John Sutton for helpful comments and suggestions.

I. INTRODUCTION

The Market in Financial Instruments Directive (MiFID) came into effect on 1 November 2007 to regulate the European financial industry. The directive intends to complete the process started with the 1993 Investment Services Directive (ISD) and provides high-level principles to foster a fair, competitive, transparent, efficient and integrated European financial market.

Overall the MiFID aims to increase competition by creating a common harmonized European market for financial products and to foster client protection through improved transparency, suitability requirements and best execution principles. In particular, it abolishes the so-called "concentration rule" that allowed, in the past, member states to impose that securities admitted to trading on a regulated market have to be traded only on regulated markets. The MiFID allows, in contrast, the provision of trading services to a variety of trading venues, namely Regulated Markets, Multilateral Trading Facilities and Systematic Internalizers.

The directive challenges therefore the market power of existing trading venues by fostering competition between alternative venues in order to achieve better execution and ultimately lower explicit costs of trading for investors (these include, in general, execution, settlement and clearing fees). As Bulow and Klemperer (2008) show, although in a different setting, *potential* competition is not a good substitute for *actual* competition. In this paper I address the question of whether fostering competition between alternative trading venues alone may or not be able to impact actual competition in the market. I consider two reasons for why it may not: (a) cash trading exhibits direct network effects and (b) there may be cases where, even though competing trading venues offer the same security, they can not be considered actual substitutes due to post-trading constraints.

In the presence of network effects, fostering venue competition is not enough to challenge the market power of existing trading venues as competitiveness depends not only on explicit but also on implicit trading costs. Implicit trading costs include, in general, the bid-ask spread, the potential impact of a trade, and the opportunity cost of missed trades and, in this setting, are important as agents prefer to place an order in a venue where a large number of other agents also place their orders. In other words, participants value liquidity and although there is no uncontroversial definition of liquidity, the negative correlation between liquidity and implicit trading costs is generally accepted. In particular, the choice of a venue with a large number of other investors translates into

lower implicit trading costs as it stabilizes the market price of a financial instrument, and reduces the extent to which placing an order has an adverse effect on the corresponding price.

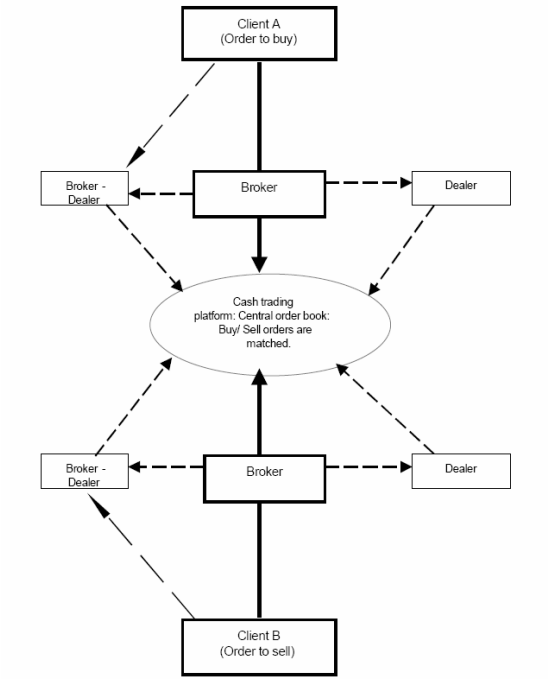
In what concerns the second reason, there are cases that even though financial agents can *a priori* choose between a set of competing trading venues to execute an order, the services offered can not actually be considered real substitutes or fungible as different trading venues may imply different settlement arrangements. If a financial agent can choose a trading venue, but can not choose the post-trading arrangements, then competition between trading venues is limited as settlement in different central securities depositories (CSD) implies higher costs for financial agents. If a financial agent wants to sell securities previously bought using trading venue A, then this venue has an advantage relatively to all other trading venues with different settlement arrangements because using different CSD necessarily implies higher trading costs. This advantage may exist even in cases where venue A may *a priori* not be offering the best price. There can not be real competition between trading venues if financial agents can not freely choose post-trading arrangements.

In this paper I propose an empirical framework to evaluate the actual degree of competition between trading venues. This empirical approach constitutes, for the best of my knowledge, one of the first attempts to structurally model financial trading, which is instrumental for measuring empirically the impact of network effects and of the bundle of trading and post-trading services as barriers to competition. This evaluation is provided in the companion paper, Ribeiro (2008).

I specify a structural discrete-choice multinomial random-coefficients logit demand model for trading following Berry, Levinsohn, and Pakes (1995) that takes into account the trade-off between explicit and implicit trading costs following Pagano (1989). The model is flexible in the sense that the implied substitution patterns do not suffer from the problem of the Independence of Irrelevant Alternatives (IIA) property characteristic of more standard multinomial logit models. Furthermore, following the demand modelling literature, the error term is structurally embedded in the model and thereby circumvents the critique provided by Brown and Walker (1989) related to the addition of add-hoc errors and their induced correlations.

The paper proceeds in eight sections. After this part of the introduction, section 2 briefly describes the trading mechanisms and the MiFID-induced changes. Section 3

FIGURE I - THE TRADING MECHANISM



Source: Pagano and Padilla (2005).

overviews the relevant literature. Section 4 presents the discrete-choice demand model and section 5 establishes some estimation and identification issues. Section 6 covers some data issues for an empirical implication. Section 7 concludes.

II. THE ECONOMICS OF TRADING AND MiFID

The process of trade starts with investors sending their buying or selling orders to a broker or a broker-dealer. If investors choose the former, the broker receives the order and can decide by one of two options: (a) can place it directly on a trading venue order book or (b) can decide to go indirectly via a dealer. If the broker opts for option (b) or the investors send their orders directly to a broker-dealer then the dealer (or broker-dealer depending on the case) can match the order from its own inventory, place the order on a trading venue or go to another dealer. The process of trading involving an electronic trading platform is illustrated in Figure I.

The paper focuses on trading venue competition and for that reason models the choice of venue to execute an order by brokers, dealers and broker-dealers (henceforth

financial agents). On this respect, MiFID promoted a significant change in the shape of the industry as it abolished the national boundaries for equity trading within the European Union as of 1 November 2007.

MiFID aims to increase competition by creating a common harmonized European market for financial products and to foster client protection through improved transparency, suitability requirements and best execution principles. In particular, it abolishes the so-called "concentration rule" that allowed, in the past, member states to impose that securities admitted to trading on a regulated market have to be traded only on regulated markets. The MiFID allows, in contrast, the provision of trading services to a variety of trading venues, namely Regulated Markets (RM), Multilateral Trading Facilities (MTF) and Systematic Internalizers (SI).

RM or MTF are entities that offer multilateral trading for financial instruments (such as an order book), with slightly different standards applying to each, whereas SI refer to financial firms which, on an organized, frequent and systematic basis, deal on own account by executing client orders outside a RM or an MTF.

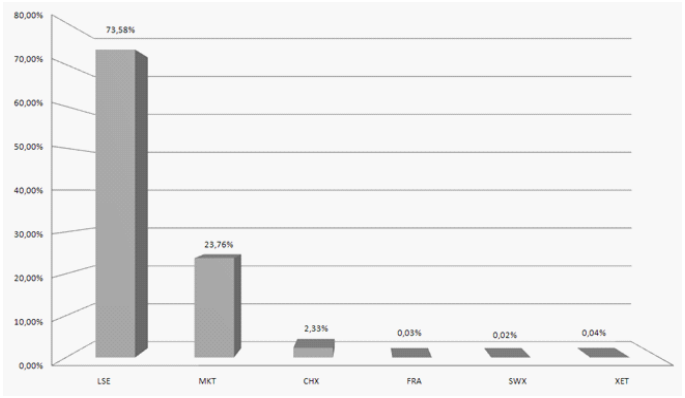
As an illustration of the MiFID-induced changes, consider a financial agent wanting to trade BP plc or TOTAL S.A. securities. The order can be routed using not only regulated markets like the London Stock Exchange, Euronext or Frankfurt Stock Exchange, but also multilateral trading facilities like Chi-X or systematic internalizers like ABN AMRO, Goldman Sachs or UBS.

Figures II present the median volume market shares for BP and TOTAL securities since November 2007. As it would be expected the larger national regulated market (LSE for BP and Euronext-Paris for TOTAL) still accounts for the majority of the traded volume, but multilateral trading facilities like Chi-X and systematic internalizers (aggregated in the figure using data from Markit BOAT) have a non-negligenciable position in the market.

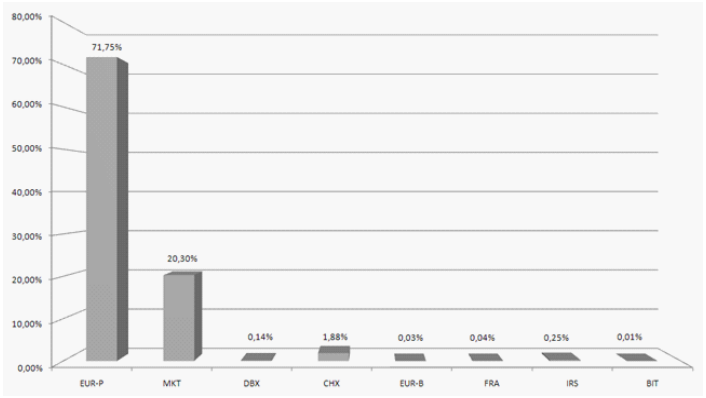
Given the set of MiFID induced alternative trading venues, financial agents have to choose the venue through which to route a certain order. An important factor to take into consideration when deciding refers to the explicit trading costs of each venue. These costs can be decomposed into costs of executing an order (trading fees) and post-trade costs (clearing and settlement fees). Clearance refers to the validation of a trade and the subsequent establishment of the obligations of the parties to the trade (what each owes and is entitled to receive). Settlement is the process during which buyer and seller

Figure II

Venue Median Market Shares for BP Security

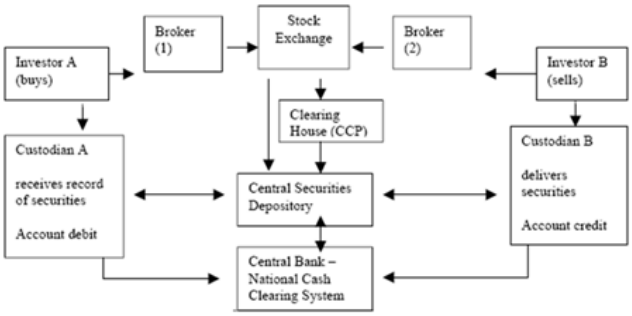


Venue Median Market Shares for TOTAL Security



Source: Author's calculations. Market shares computed for traded volume.

Figure III - Clearing and Settlement Flows



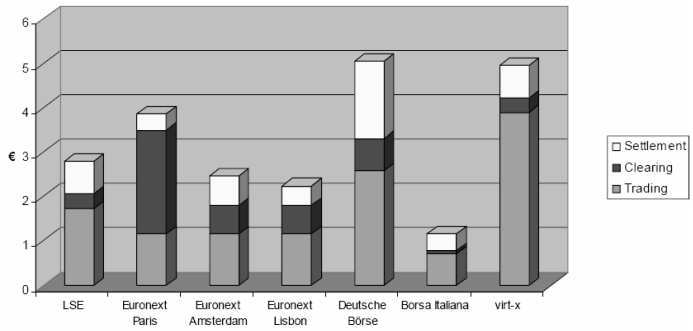
Source: Carvalho (2004).

details are matched and the security changes ownership against the appropriate payment. Clearing and settlement services are typically performed by specializing institutions: the transfer of ownership is carried out by a central securities depository or an international central securities depository, whereas the banking/payment system handles the payment of funds. Figure III present the flows involved in the clearing and settlement of a trade.

Figure IV show the explicit trading costs (and the respective decomposition) faced by a typical financial agent and it is clear that those vary substantially across trading venues, not only in absolute terms but also in their composition. The analysis of the figure may suggest an intriguing question: given that competing venues have different explicit trading costs, what prevents trade to concentrate on the venue which offers the lowest fees? The justification is two-folded. First, explicit trading costs are typically (although not always) a function of an agents' trading profile. What this implies is that the different fees schedules can be such that venue A and venue B can coexist because they attract agents with diverse trading profiles. Figure IV presents the explicit trading costs solely for the typical financial agent. Second, the comparison of the relative explicit trading costs of each venue is not the only criteria upon which financial agents base their decisions. They also take into consideration the implicit trading costs, which typically include the bid-ask spread, the potential impact of a trade, and the opportunity cost of missed trades.

The importance of the implicit trading costs arises because cash trading exhibits direct network effects. The valuation of a venue by financial agents increases the more that other agents choose the same venue. In other words, agents prefer to place an order in a venue where a large number of other agents also place their order since this reduces the costs of finding a counterparty, which in turn may increase the valuation of

FIGURE IV - DECOMPOSITION OF EXPLICIT COSTS PER TRADE



Source: European Commission (2006). Data refers to 2004.

that venue even more. In particular, the choice of a venue with a large number of other investors translates into lower implicit trading costs as it (a) stabilizes the market price of a financial instrument, and (b) reduces the extent to which placing an order has an adverse effect on the corresponding price.

Pagano (1989) shows that, within a given type of trading profile, if the explicit trading costs are equal across venues, the direct network effects promote the concentration of trade on only one venue. However, if the low explicit trading costs of a venue are traded-off against higher implicit trading costs, multiple trading venues can coexist in equilibrium, even within agents with similar trading profiles.

In the presence of network effects, fostering venue competition is therefore not enough to challenge the market power of existing trading venues as competitiveness depends not only on explicit but also on implicit trading costs. In fact underestimating the importance of network effects can often lead to a dismal failure similar. As an illustration consider the case of Jiway, a pan-European trading platform for retail investors launched in the last quarter of 2000 by Morgan Stanley and the Swedish company OM. The two companies invested $100 million on the project that promised access to 6,000 European securities, but it turned out to be unable to attract liquidity: in January 2001 it executed 1,996 trades, in February 474 trades, and in March 577 trades. As a result, by the end of 2002, Jiway was shut down.

Another illustration is provided by Chi-X, a multilateral trading facility set up in the first quarter of 2007. Chi-X soon understood that if it wanted to successfully attract trades it needed to balance the high implicit trading costs (due to the low liquidity) with extremely low explicit trading costs. The solution (up to this moment with very

optimistic results) has been to offer a fee schedule that reverses the standard in the industry and includes, in certain cases, a negative execution fee that translates into a payment from the venue to the agent.

Competition between alternative trading venues may thereby be limited due to direct network effects. However those effects do not constitute the only barrier to venue competition. The bundling of trading and of post-trading services constitute another barrier. The reason is that even though financial agents can *a priori* choose between a set of competing trading venues to execute an order, the services offered can not actually be considered real substitutes or fungible as different trading venues may imply different settlement arrangements.

Consider, as an illustration, a financial agent with an order to trade Royal Dutch securities. The agent can execute the order on a set of alternative venues from Euronext Amsterdam to Deutsche Borse. However because post-trading services are typically bundled with trading services, when the agent chooses a venue, she is implicit choosing also the corresponding post-trading provider. Table I presents the trading venues and associated central securities depositories for Royal Dutch securities.

Table I - Royal Dutch Trading and Pos-Trading (Venue/CSD)

Euronext Amsterdam / Euroclear Amsterdam
London Stock Exchange / Euroclear Amsterdam
Chi-X / Euroclear Amsterdam
Virt-X / Euroclear Bank
Deutsche Borse / Clearstream Banking Frankfurt

Source: Misra (2007).

In the illustration above, only the securities trading in Euronext Amsterdam, London Stock Exchange and Chi-X are fully fungible as they settle in the same CSD, Euroclear Amsterdam. Trading Royal Dutch in Virt-X or Deutsche Borse may imply settlements across different CSD with associated higher costs. Carvalho (2004) concludes that the costs of clearing and settlement across different CSD within Europe are 42% higher than if using the same CSD. As a result, venues that settle in the same CSD have an advantage when compared with those that settle in different CSD. This advantage may exist even in cases where one venue may *a priori* not be offering the best price. In sum, there can not be real competition between trading venues if financial agents can not freely choose post-trading arrangements.

In the discussion above, I present arguments that sustain that barriers to venue competition may exist even after MiFID. As a last note, I would like to point that if actual competition can have a extremely positive effect, it may also have a negative one: a fragmentation effect. When different trading venues coexist, markets become fragmented and the liquidity available in any one setting is reduced, thereby potentially limiting any market's ability to provide stable prices. The bid-ask spreads might be greater and daily securities returns might have a larger variance. Moreover, as liquidity facilitates the crucial price discovery role of markets, as order flow fragments, the ability of prices to aggregate information can be reduced, and with it the efficiency of the market.

MiFID addresses this point by requiring every venue not only to publish the price, volume and time of a transaction as close to real-time as possible, but also to do it in a way that is easily accessible to other market participants. Furthermore, it also consolidates the hitherto fragmented market of European over-the-counter (OTC) securities. For these reasons, the fragmentation issues of increased trading venue competition may be less significant for MiFID.

In this paper, I propose an empirical framework to evaluate the effective competition between alternative trading venues.

III. RELEVANT LITERATURE

The literature on market dominance begins with Gilbert and Newbery (1982) and Reinganum (1983) who show that a monopolist can maintain her dominance due to stronger incentives for preemptive innovation. Other contributions include Budd, Harris and Vickers (1993), Cabral and Riordan (1984), Athey and Schmutzler (2001) and Cabral (2002). Budd, Harris and Vickers (1993) analyze the dynamics of market structure in a duopoly and, in particular, in what circumstances we may see a process of increasing dominance sourced on higher levels of technology. Cabral and Riordan (1984) investigate another source of eventual market dominance, the hypothesis that due to a learning curve, unit costs may decline with cumulative production. Athey and Schmutzler (2001) model an oligopolistic setting to examine conditions under which dominance sourced in ongoing investment may emerge. Cabral (2002) considers a similar setting but where firms choose the amount of resources to invest and how to allocate those resources.

This paper analyzes market dominance sourced on (a) network effects and (b) trading and post-trading bundling. The literature on network effects begins with Katz and Shapiro (1985) and from then on it has developed along two different directions. Katz and Shapiro (1994), Economides (1996), Shy (2001), and Farrell and Klemperer (2006) provide an excellent overview of this literature. One of the strands of the literature tries to empirically measure the effect of network effects, whereas the other studies its implications. In what concerns the second source of market dominance, competition between trading and post-trading services has been modelled by Tapking and Yang (2004) and Koppl and Monnet (2003). The former studies different forms of industry structures between venues and post-trading firms, whereas the latter analyzes the impact of integrating the two services.

A number of papers have explicitly studied venue competition. The seminal work is from Hamilton (1979) who establishes the two opposite effects of multi-venue trading and reports empirical estimates of the effect of off-boarding trading on liquidity and volatility of NYSE stocks. Multi-venue trading promotes lower explicit trading costs via higher competition but also has a fragmentation effect. When different trading venues coexist, markets become fragmented and the liquidity available in any one setting is reduced, thereby potentially limiting any market's ability to provide stable prices. The bid-ask spreads might be greater and daily securities returns might have a larger variance. Moreover, as liquidity facilitates the crucial price discovery role of markets, as order flow fragments, the ability of prices to aggregate information can be reduced, and with it the efficiency of the market. Hamilton finds that the competitive effect exceeds the fragmentation effect, and that both effects are small.

In general, followers of Hamilton's legacy use a reduced-form strategy that regress spreads and liquidity on stock and market characteristics that include a competition variable. More recent examples include Weston (2002) and Gresse (2006). Weston (2002) investigates whether the shift towards electronic communication networks leads to tighter bid-ask spreads and greater depths. He finds that this particular competition has a significant negative impact on bid-ask spreads, but no significant impact on quoted depth. Gresse (2006) studies the impact of crossing networks on the liquidity of the dealer market segment of the London Stock Exchange (SEAQ). She finds that spreads decrease due to competition but no fragmentation effect is detected.

In parallel to the above approach, the literature has also evolved towards more structural and micro-founded strategies of modelling financial markets of which Hortaçsu and

Syverson (2004) and Cantillon and Ying (2007) are some recent examples. Hortaçsu and Syverson (2004) investigate the role that nonportfolio fund differentiation and information/search frictions play in creating two salient features of the mutual fund industry: the large number of funds and the sizable dispersion in fund fees. Cantillon and Ying (2007) study the determinants of the dynamics of the market for the future on the Bund.

I propose to estimate a structural discrete-choice demand model for trading following Berry, Levinsohn, and Pakes (1995) that tries to reconcile the advantages of Hamilton (1979)'s approach with the desirable features of a micro-founded model, taking into account two eventual barriers to competition, network effects as well as the bundle of trading and post-trading services.

IV. DEMAND FOR TRADING

The trading decision can be decomposed in two stages. First, investors decide the order characteristics and send it to an financial agents to be executed. Second, after receiving the order the agent decides the trading venue where to execute it, conditional on the order characteristics received. In this paper, I take the first stage as given and propose to model the second stage choice by financial agents. An interesting and natural extension will be the incorporation of the first-stage into the model framework.

Consider that in period $t = 1, \ldots, T$ an investor sends an order with characteristics k (including e.g. the code of the security, the direction and the volume to be traded) to financial agent $i = 1, \ldots, I$ for her to execute. After receiving the order, the financial agent has choose the trading venue where to execute the order subject to her internal best execution policy that had, under MiFID, to have been previously accepted by the investor.

The best execution policy defines the agent's commitment towards the investor to achieve the best possible result for their clients taking into account price, costs, speed, likelihood of execution and settlement, size, nature or any other consideration relevant to the execution of the order. An alternative view for the agent's best execution policy is to think of it as an auction where the agent allocates the order across the alternative trading venues according to an allocation rule known to the investor but unknown to the econometrician.

I propose to estimate the allocation rule by specifying a structural multinomial random-coefficients logit discrete-choice demand model for trading following Berry, Levinsohn, and Pakes (1995) where in each period t heterogeneous financial agents i consider to execute an order with characteristics k in a trading venue $v = 0, 1, \ldots, V$, where $v = 0$ denotes the outside option of executing the order over-the-counter.

I model financial agents as making myopic decisions or equivalently to have static expectations about the future based on the fact that the best execution policy has to be applied on a trade by trade case. I therefore assume the conditional indirect utility that financial agent i obtains from executing an order of characteristics k at venue v in period t to be of the form

$$u_{ikvt}\left(p^{e}_{ikvt}, w_{vkt}; \theta_i\right) = u^{*}\left(p^{e}_{ikvt}, w_{kvt}; \theta_i\right) + \varepsilon_{ikvt}, \tag{1}$$

where w_{kvt} represents a vector of attributes for the order, venue and time period, and p^{e}_{ikvt} denotes the expected all-in explicit trading costs faced by the financial agent, which include execution, clearing and settlement fees. Because the fees schedules are typically a function of agent i's trading profile[1] during a certain time period as well as of subset of order characteristics, the explicit trading costs p^{e}_{ikvt} are unknown ex-ante and are indexed by i and k. In order to explicitly illustrate the non-linearity of the fees schedules, I will denote $p^{e}_{ikvt} = p_{vt}\left(z^{e}_{i}, k\right)$, where z^{e}_{i} expresses the expectation of agent i's trading profile. Lastly, financial agents heterogeneity in their allocation rule for trading venues enters the conditional indirect utility through agent-specific valuation θ_i of the different elements included in the best execution policy and an additive preference shock ε_{ikvt}.

Among the attributes of a trading venue, w_{kvt}, that impact the choice of agents are the implicit trading costs b_{kvt} as cash trading exhibits network effects and participants value liquidity. Although there is no uncontroversial definition of liquidity, the negative correlation between liquidity and implicit trading costs is generally accepted. A large installed base of agents trading at venue v promotes lower implicit trading costs as it (a) stabilizes the market price of a security, and (b) reduces the extent to which placing an order has an adverse effect on the corresponding price. As a side note, these network effects can be artificially reinforced by fees schedules that are decreasing in trade volume.

Following Davis (2006) and Chen et al. (2007), $u^{*}\left(\cdot\right)$ is assumed to be of the form

$$u^{*}\left(p^{e}_{ikvt}, w_{kvt}; \theta_i\right) = -p_{vt}\left(z^{e}_{i}, k\right) - \gamma_i g\left(b_{kvt}\right) + x^{'}_{kvt}\beta_i + \xi_{kvt} + \tau_k + \eta_t, \tag{2}$$

[1] Volume discounts can reflect venue economies of scale that are passed to agents.

where:

(*a*) the vector of characteristics w_{kvt} is split between the implicit trading costs b_{kvt}, a K-dimensional vector of observables, x_{kvt}, and a vector of unobserved (to the econometrician) characteristics, whose mean valuation for orders with characteristics k executed in venue v in period t across financial agents is given by ξ_{vt};

(*b*) The increasing function $\gamma_i g(\cdot)$ captures the network effects, where $\gamma_i \geq 0$ is the parameter that controls the strength of those network effects.

(*c*) τ_k and η_t denote an order and time fixed effect, respectively; and

(*d*) θ_i denotes the parameters of estimation: $\theta_i = (\gamma_i, \beta_i)'$. I normalize the valuation of all the different elements in the allocation rule with reference to the valuation of the explicit trading costs.

For completeness, the financial agent can also choose to execute the order over-the-counter. The conditional indirect utility from the outside option is assumed to be $u_{ik0t} = \xi_{k0t} + \varepsilon_{ik0t}$. Following the literature, I will normalize without loss of generality $\xi_{k0t} = 0$ as due to the ordinality of utility, only $\xi_{kvt} - \xi_{k0t}$ matters for the agent's choice of venue.

The parameters of estimation γ_i^* and β_i^* are indexed by agent in order to capture the fact that the valuations of the different elements in the allocation rule can depend on agents's characteristics. In particular, I will allow those parameters to be a function of the expectation of the agents' trading profiles z_i^e

$$\begin{pmatrix} \gamma_i^* \\ \beta_i^* \end{pmatrix} = \begin{pmatrix} \gamma \\ \beta \end{pmatrix} + \theta^o z_i^e, \tag{3}$$

where θ^o denote coefficients that will express the heterogeneity of agents in reference with their trading profile. As a result the parameters to be estimated reduce to $\theta = (\gamma, \beta, \theta^o)'$

Among the characteristics k of an order is obviously the code of the security to be traded. Let the index j denote the identity of that security and k' index the remaining characteristics of the order. After substituting equation (3) into the conditional indirect utility function (1), it is possible to summarize the financial agent's conditional indirect utility as a sum of two terms: a first term that depends only on the identity of the

security and is common across agents , $\delta_{jvt} = -\gamma g\left(b_{jvt}\right) + x_{jvt}^{'}\beta_1 + \xi_{jvt} + \tau_j + \eta_t$, and a second term, $\mu_{ikvt} + \varepsilon_{ikvt}$, that introduces agent heterogeneity

$$u_{ikvt} = \delta_{jvt} + \mu_{ikvt} + \varepsilon_{ikvt}, \tag{4}$$

where

$$\mu_{ikvt} = x_{k'vt}^{'}\beta_2 - p_{vt}\left(z_i^e, k\right) + \left[g\left(b_{jvt}\right), x_{kvt}^{'}\right]\theta^o z_i^e.$$

Given the heterogeneity of the financial agents specified in the model, the solution to the maximization problem of the indirect conditional utility over all the different venues will vary from one agent to another, depending on their specific attributes $(z_i^e, \varepsilon_{ikt})$ where $\varepsilon_{ikt} = (\varepsilon_{ik0t}, \ldots, \varepsilon_{ikVt})$. As a result, conditional on the order characteristics, the set of financial agents that execute the order to trade at venue v in period t is then

$$A_{kvt}\left(x_t, p_t, \delta_t; \theta\right) = \left\{\left(z_i, \varepsilon_{ik0t}, \ldots, \varepsilon_{ikVt}\right) | u_{ikvt} > u_{ikgt} \forall g \text{ s.t. } v \neq g\right\}, \tag{5}$$

where x_t, p_t and δ_t are the vectors of observed characteristics, explicit trading costs and deltas. If the preference shock ε_{ikvt} follows an i.i.d. extreme value distribution, the probability that agent i opts for venue v to execute order with characteristics k in period t is then given by the following multinomial logit type expression

$$P_{ikvt}\left(x_t, p_t, \delta_t; \theta, k\right) = \frac{e^{\delta_{jvt} + \mu_{ikvt}}}{1 + \sum_q e^{\delta_{jqt} + \mu_{ikqt}}}. \tag{6}$$

Integrating over the distribution of agents' specific attributes and order characteristics (z_i, k) yields market-level share for venue v in each period t

$$s_{jvt}\left(x_t, p_t, \delta_t; \theta\right) = \int_{A_{vt}} \frac{e^{\delta_{jvt} + \mu_{ikvt}}}{1 + \sum_q e^{\delta_{jqt} + \mu_{ikqt}}} dP^*\left(z, k\right), \tag{7}$$

where $P^*\left(z, k\right)$ denotes the population joint distribution function of the agent types and order characteristics (z_i, k), not necessarily independent.

V. IDENTIFICATION AND ESTIMATION PROCEDURE

I now proceed with a description of the procedure to estimate the parameter vector $\theta = (\gamma, \beta, \theta^o)'$. The data available to the researcher is crucial for the estimation procedure.

In what follows, I will assume that a know joint distribution of the agent types and order characteristics is available. However, the procedure can easily be modified for the case where that distribution is unavailable and one distribution needs to be assumed, incorporating into the utility specification its unknown parameters, to be estimated jointly with the other parameters of the model.

The estimation algorithm encompasses four steps that I now describe.

Step One Set initial values for the mean utilities, δ_t, and for the parameters of estimation, θ.

Step Two Approximate the predicted market-level shares. The key difficulty with the random-coefficients multinomial logit model has to do with the fact that no closed form expression exists for the integral that defines those predicted shares

$$s_{jvt}\left(x_t, p_t, \delta_t; \theta\right) = \int_{A_{vt}} \frac{e^{\delta_{jvt}+\mu_{ikvt}}}{1+\sum_q e^{\delta_{jqt}+\mu_{ikqt}}} dP^*\left(z,k\right). \tag{8}$$

As the computation of the above expression is, in general, problematic, the literature follows Pakes (1986), Pakes and Pollard (1989), and McFadden (1989) and approximates that intractable integral by a simulation estimator. In what the particular choice of the simulation estimator is concerned, the smooth simulator has been the prevailing approach. To compute it, ns pseudo-random vectors of unobserved agent attributes $(z_1^r, \ldots, z_{ns}^r)$ and order characteristics $(k_1^r, \ldots, k_{ns}^r)$ are drawn from $dP^*\left(z,k\right)$, and, given the initial values of δ_t and θ, used to obtain $\delta_{kvt} + \mu_{ikvt}^r$ where

$$\mu_{ikvt}^r = x_{k'vt}^{'}\beta_2 - p_{vt}\left(z_i^r, k^r\right) + \left[g\left(b_{kvt}\right), x_{kvt}^{'}\right]\theta^o z_i^r. \tag{9}$$

The smooth estimator that simulates the aggregate market shares is, then, given by

$$s_{jvt}\left(x_t, p_t, \delta_t; \theta, P^{ns}\right) = \frac{1}{ns}\sum_{i=1}^{ns} \frac{e^{\delta_{jvt}+\mu_{ikvt}^r}}{1+\sum_q e^{\delta_{kqt}+\mu_{ikqt}^r}}, \tag{10}$$

where P^{ns} denotes the empirical distribution of the simulation draws. Please note that this estimator, in contrast with other simulation estimators[2], by integrating the ε's

[2]Please see Berry, Levinsohn, and Pakes (1995) for a detailed survey on the optimal importance sampling simulator, and the appendix to Nevo (2000) for an analysis on the naive frequency estimator.

analytically, circumvents the need to draw them and, consequently, limits the simulation error to the sampling process. It is also instrumental in obtaining simulated market-level shares that are smooth functions, positive and sum to one.

As a final note I would like to stress, as Berry, Linton, and Pakes (2004) point out, that the introduction of simulation error influences the asymptotic distribution of the estimator and, therefore, needs to be explicitly taken it account. On this subject please see step four below.

Step Three Estimate the econometric error, ξ_{jvt}, as a function of the parameters of estimation θ. The mean utility δ_{jvt} can not be solved for analytically, but Berry, Levinsohn, and Pakes (1995) proved that, for a given θ, the mapping of values of δ_{jvt} into themselves is a contraction mapping with modulus less than one, and therefore that it is possible to solve for the unique δ_{kvt} that matches the simulated market-level shares, $s_{jvt}\left(x, p_t, \delta_t; \theta, P^{ns}\right)$ with the observed ones, s^n_{jvt}, for all j, v and t, recursively,

$$\delta^k_{jvt}\left(\theta\right) = \delta^{k-1}_{jvt}\left(\theta\right) + \ln\left[s^n_{jvt}\right] - \ln\left[s_{jvt}\left(x_t, p_t, \delta^{k-1}_t; \theta, P^{ns}\right)\right], \tag{11}$$

as the iterations converge geometrically to the unique fixed point, where the simulated market-level shares $s_{jvt}\left(x_t, p_t, \delta_t; \theta, P^{ns}\right)$ have to be computed at every new iterated δ^k_t. Denote the fixed point by $\delta_{jvt}\left(s^n_t, \theta, P^{ns}\right)$ where s^n_t represents the vector of observed aggregate market shares.

Given the unique fixed point, it is relatively straithforward to obtain an estimate of the econometric error as a function of the data, x, p_t, s_t, the parameters of estimation, θ, and the simulation process, P^{ns},

$$\xi_{jvt}\left(s^n_t, \theta, P^{ns}\right) = \delta_{jvt}\left(s^n_t, \theta, P^{ns}\right) + \gamma g\left(b_{jvt}\right) - x^{'}_{jvt}\beta_1 - \tau_j - \eta_t. \tag{12}$$

Step Four Estimate the parameters θ. Typically, the estimation procedure relies on an identifying restriction over the distribution of the true econometric error, obtained by evaluating equation (12) at $n = ns = \infty$, that is, $\xi_{jvt}\left(s^\infty_t, \theta, P^\infty\right)$.

An econometric issue with the above estimation procedure relates to an eventual correlation between trading costs and the econometric error term. This correlation is expected as trading costs typically incorporate some information that the econometrician does not possess and, thereby, has to include in the econometric error term. Due to this

eventual correlation, instrumental variables techniques are, therefore, required. I assume, however, as it is standard in the literature, the unobserved characteristics to be mean independent of the observed ones (please see Berry, 1994).

I follow Davis (2006) and aim to identify the parameters of the model by applying GMM to two sets of population moment conditions,

$$E\left[\bar{\xi}_{jv}\left(s_t^{\infty},\theta^*,P^{\infty}\right)|Z_{jv}\right]=0 \text{ and } E\left[\xi_{jvt}\left(s_t^{\infty},\theta^*,P^{\infty}\right)|\right]=0 \text{ for } t=2,\ldots,T, \quad (13)$$

where ξ_{jvt} denotes the unobserved (to the econometrician) valuation of security j at venue v in period t and $\bar{\xi}_{jv}$ denotes its average across time. The first set of moment conditions restrict $\bar{\xi}_{jv}$ to be uncorrelated with a set of instruments $Z_{jvt}=\left[z_{jvt}^1,\ldots,z_{jvt}^M\right]$ at the true parameter values θ^*. $\bar{\xi}_{jv}$ is used so that the GMM standard errors provide a conservative estimate of the amount of information in the sample, since ξ_{jvt} typically exhibits significant positive correlation. The second set of moments identifies the $(T-1)$ period fixed effects, η_t, that capture the strong within-time period seasonality observed in the data.

The above population moment conditions can be used, akin to Hansen (1982), to render a method of moments estimator of θ^*, by interacting the estimated econometric error with the set of instruments, and search for the value of the parameters, θ, that set the sample analogues of the moment conditions as closed as possible to zero. Let $G_{n,ns}\left(\theta\right)$ denote the sample analogues of the moment conditions,

$$G_{n,ns}\left(\theta\right)=\frac{1}{n}\sum_{t=1}^{T}\sum_{v=1}^{V}\sum_{j=1}^{J^{vt}}\begin{pmatrix}\bar{\xi}_{jv}\left(s_t^n,\theta,P^{ns}\right)Z_{jvt}\\ \xi_{jvt}\left(s_t^n,\theta,P^{ns}\right)\end{pmatrix}=\frac{1}{n}\sum_{t,v,j}\psi\left(\theta\right). \quad (14)$$

Formally, the method of moments estimator, $\hat{\theta}$, is therefore the argument that minimizes the weighted norm criterion of $G_{n,ns}\left(\theta\right)$, for some weighting matrix A_n with rank at least equal to the dimension of θ,

$$\hat{\theta}=\arg\min_{\theta}\quad\left\|G_{n,ns}\left(\theta\right)\right\|_{A_n}=G_{n,ns}\left(\theta\right)'A_nG_{n,ns}\left(\theta\right). \quad (15)$$

The strong non-linearity of the objective function requires a minimization routine. The standard practice in the literature has been to use either the Nelder-Mead (1965) nonderivative "simplex" search method or a quasi-Newton method with an analytic

gradient (see Press at al., 1994). The latter has the important (computational) advantage of being two orders of magnitude faster than the former. However, because the first method is more robust and less sensitive to starting values, I will perform the search using the Nelder-Mead (1965) nonderivative "simplex" search.

The non-linear search over θ can be simplified by making use of the fact that the first order conditions for a minimum of $\|G_{n,ns}(\theta)\|_{A_n}$ are linear for the subset $\theta_1 = (\gamma, \beta)$ of the parameters of estimation, $\theta = (\theta_1, \theta^o)$. Consequently, it is possible, given the standard instrumental variables results, to express θ_1 as function of θ^o, and limit the non-linear search over θ^0,

$$\theta_1 = \left(Q'ZA_n^{-1}Z'Q\right)^{-1} Q'ZA_n^{-1}Z'\delta\left(\theta^0\right). \tag{16}$$

where Q denotes the matrix of trading costs and observed characteristics, Z denotes the matrix of instruments, and, finally, δ denotes the matrix of mean utilities, expressed only in terms of θ^0 after concentrating out θ_1.

In what the choice of instruments is concerned, I follow Berry, Levinsohn, and Pakes(1995) and suggest using rival characteristics as instruments, since we would expect variations of a given security/venue's trading costs to be correlated to variations in the characteristics of competing products.

VI. EMPIRICAL APPLICATION

The procedure described in the previous sections relies on the availability of market-level data on trading costs, observed characteristics and market shares of the different alternative trading venues.

As mentioned already, trading costs can be explicit and implicit. The explicit costs include, in general, execution, settlement and clearing fees, whereas the implicit costs include, in particular, the bid-ask spread, the potential impact of a trade, and the opportunity cost of missed trades.

Information on execution, settlement and clearing fees can be obtained via the published fee schedules for the different venues. Given that typically (although not always) those fee schedules are a function of agent's i trading profile, the solution may follow

European Commission (2006) and define four representative trading profiles: eg. "typical volume and value trades' agent" *vs* "large volume of low value trades' agent" *vs* "large volume of high value trades' agent" *vs* "small volume of low value trades' agent". Furthermore, as trading profiles are typically only known ex-post, explicit trading costs may be expressed as a function of a weighted average of past trading profiles.

A robustness check towards the explicit trading costs computed as outlined here can be provided via a no-arbitrage condition. If trading costs were of the same magnitude across venues, two securities trading in different venues would trade at the exact same price. Otherwise, opportunities for profitable arbitrage would exist and investment firms will take advantage of them until they cease to exist. With trading costs differentiated across venues, the same principle must apply up to the difference in those trading costs. Using this no-arbitrage condition, it is then possible to compute all-in explicit trading costs for the different trading venues up to a normalization. Garvey and Murphy (2006) provide evidence for a no-arbitrage condition across venues for Nasdaq-listed stocks. They analyze 20 Nasdaq stocks and find that during only 0.5% of the trading time there existed arbitrage opportunities, lasting on average 7 seconds.

In what the implicit trading costs is concerned, the following variables may be computed for each stock: the effective percentage spread, stock volatility and trading turnover. I suggest following Stoll (2000) and Jain (2001) and aggregate the different variables at a weekly or monthly frequency as it reduces substantially measurement errors due to random daily fluctuations.

The effective percentage spread is a measure of trading costs and is defined as the difference between the transaction price and the current mid-quote for time period t,

$$EPS_{jt} = mean\left[2\frac{|PR_{jd} - M_{jd}|}{PR_{jd}}\right], \tag{17}$$

where M_{jd} is the quote mid-point, i.e. $(A_{jd} + B_{jd})/2$, A_{jd} denotes the ask price, B_{jd} the bid price, PR_{jd} the effective transaction price of instrument j in day d, and $mean\,[\cdot]$ represents the average taken over the days included in period t. This measure takes into account the fact that trades can occur either inside or outside the quoted spread. Therefore, it incorporates both the impacts of market spreads and market impact on trading costs, even if it does not allow the separation of the two effects.

The stock volatility is defined, following Ding and Charoenwong (2003), as the standard deviation over the average of the quoted mid-point within each time period,

$$SV_{jt} = \frac{sd\left[M_{jd}\right]}{mean\left[M_{jd}\right]}, \tag{18}$$

where $sd\left[\cdot\right]$ represents the standard deviation taken over the days included in period t.

Lastly, the trading turnover is defined as the ratio between monetary trading volume and market capitalization,

$$TT_{jt} = mean\left[\frac{TV_{jd}}{MC_{jd}}\right], \tag{19}$$

VII. CONCLUDING REMARKS

The Market in Financial Instruments Directive (MiFID) aims to increase competition and to foster client protection in the European financial market. Among other provisions, it abolishes the concentration rule and challenges the market power of existing trading venues.

The directive introduces venue competition in order to achieve better execution and ultimately lower costs of trading. However, the fostering venue competition may not be enough. In this paper I address the question of whether fostering competition between alternative trading venues alone may or not be able to impact actual competition in the market. I consider two reasons for why it may not: cash trading exhibits direct network effects and trading and post-trading bundling.

In this paper I propose an empirical framework to evaluate the actual degree of competition between trading venues. This empirical approach constitutes, for the best of my knowledge, one of the first attempts to structurally model financial trading, which is instrumental for measuring empirically the impact of network effects and of the bundle of trading and post-trading services as barriers to competition. This evaluation is provided in the companion paper, Ribeiro (2008).

VIII. REFERENCES

Anolli, M., 2007, "The Impact of MiFID on the European Securities Industry: A Simulation of the Internalizer Behavior on the Italian Stock Market," *Catholic University.*

Athey, S. and A. Schmutzler, 2001, "Investment and Market Dominance," *RAND Journal of Economics*, 32, 1-26.

Avgouleas, E. and S. Degiannakis, 2005, "The Impact of the EC Financial Instruments Markets Directive on the Trading Volume of EU Equity Markets," *Athens University.*

Berry, S., 1994, "Estimating Discrete-Choice Models of Product Differentiation," *Rand Journal of Economics*, 25, 242-262.

Berry, S., J. Levinsohn, and A. Pakes, 1995, "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841-890.

Berry, S., O.B. Linton, and A. Pakes, 2002, "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," *mimeo.*

Brown, B. W. and M. B. Walker, 1989, "The Random Utility Hypothesis and Inference in Demand Systems," *Econometrica*, 57, 815-829.

Budd, C., C. Harris and J. Vickers, 1993, "A Model of the Evolution of Duopoly: Does the Asymmetry Between Firms Tend to Increase or Decrease?," *Review of Economic Studies*, 60, 543-573.

Bulow, J. and P. Klemperer, 2008, "Why do Sellers (usually) prefer Auctions," *mimeo.*

Cabral, L. M. B., 2002, "Increasing Dominance with no Efficiency Effect," *Journal of Economic Theory*, 102, 471-479.

Cabral, L. M. B. and M. H. Riordan, 1994, "The Learning Curve, Market Dominance and Predatory Pricing," *Econometrica*, 62, 1115-1140.

Cantillon, E. and P. Yin, 2007, "How and When do Markets Tip? Lessons from the Battle of the Bund," *ECB Working Paper Series*, 766.

Chen, J, U. Doraszelski, and J. Harrington Jr., 2007, "Avoiding Market Dominance: Product Compatibility in Markets with Network Effects," *mimeo.*

Davis, P., 2006, "Spatial Competition in Retail Markets: Movie Theaters," *RAND Journal of Economics*, 37, 964-982.

Ding, D. and C. Charoenwong, 2003, "Bid-Ask Spreads, Volatility, Quote Revisions, and Trades of Thinly Traded Futures Contracts," *The Journal of Futures Markets*, 23, 455–486.

Economides, N., 1996, "The Economics of Networks," *International Journal of Industrial Organization*, 14, 673-699.

EUROPEAN COMMISSION, 2006, "Competition in EU Securities Trading and post-Trading," *Working Paper of the Commission Financial Services.*

FARRELL, F. and P. KLEMPERER, 2006, "Coordination and Lock-in: Competition with Switching Costs and Network Effects," forthcoming in M. Amstrong and R. Porter (Eds), *Handbook of Industrial Organization,* Vol 3, North-Holland.

GARVEY, R. and A. MURPHY, 2006, "Crossed Markets: Arbitrage Opportunities in Nasdaq Stocks," *The Journal of Alternative Investments*, 23, 46–58.

GILBERT, R. J. and D. M. G. NEWBERY, 1982, "Preemptive Patenting and the Persistance of Monopoly Power," *American Economic Review*, 72, 514–526.

GOOLSBEE, A. and A. PETRIN, 2004, "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 72, 351–381.

GRESSE, C. , 2006, "The Effect of Crossing-Network Trading on Dealer Market's Bid-Ask Spreads," *European Financial Management*, 12, 143-160.

HAMILTON, J. L., 1979, "Market Fragmentation, Competition, and the Efficiency of the Stock Exchange," *The Journal of Finance*, 34, 171-187.

HANSEN, L. P., 1982, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

HORTACSU, A. and C. SYVERSON, 2004, "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," *The Quarterly Journal of Economics*, 119, 403-456.

JPMORGAN, 2006, "MiFID Report II", *European Equity Research.*

JAIN, P., 2001, "Institutional Design and Liquidity at Stock Exchanges around the World", *Kelley School of Business – Indiana University.*

KATZ, M.L. and C. SHAPIRO, 1985, "Network Externalities, Competition and Compability," *American Economic Review*, 75, 424-440.

KATZ, M.L. and C. SHAPIRO, 1994, "Systems Competition and Network Effects," *Journal of Economic Perspectives*, 8, 93-115.

KOPPL, T. and C. MONNET, 2003, "Guess What: It's the Settlements!," *European Central Bank discussion paper.*

MCFADDEN, D., 1989, "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 119, 403-456.

NELDER, J.A., and R. MEAD, 1965, "A Simplex Method for Function Minimization," *Computer Journal*, 7, 308-313.

NEVO, A., 2000, "A Practioner's Guide to Estimation of Random-Coefficients Logit Models of Demand," *Journal of Economics & Management Strategy*, 9, 513-548.

PAGANO, M., 1989, "Trading Volume and Asset Liquidity," *The Quarterly Journal of Economics*, 104, 255-274.

— and A.J. PADILLA, 2005, "The Economics of Cash Trading: An Overview," *LECG Report for Euronext.*

PAKES, A., 1986, "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755-784.

— and D. POLLARD, 1989, "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1057.

PRESS, W.H., S.A. TEUKOLSKY, W.T. VETTERLING, and B.P. FLANNERY, 1994, *Numerical Recipes in C*, Cambridge University Press.

REINGANUM, J. F., 1982, "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, 73, 741-748.

RYSMAN, M., 2007, "An Empirical Analysis of Payment Card Usage," *The Journal of Industrial Economics*, LV, 1-36.

RIBEIRO, R., 2008, "Market Dominance and Barriers to Competition in Venue Trading Competition," *London School of Economics.*

SHY, O., 2001, *The Economics of Network Industries* (Cambridge: Cambridge University Press).

STOLL, H., 2000, "Friction", *Vanderbilt University Working Paper*, 00-06.

TAPKING, J., and J. Yang, 2003, "Horizontal and Vertical Integration in Securities Trading and Settlement", *Bank of England working paper.*

WESTON, J., 2002, "Electronic Communication Networks and Liquidity on the Nasdaq," *Journal of Financial Services Research*, 22, 125-139.

Analyzing the Relationship between Regulation and Investment in the Telecom Sector

Hans Friederiszick (ESMT)

Michal Grajek (ESMT)*

Lars-Hendrik Röller (ESMT)

March 2008

Abstract

This study analyses the relationship between entry regulation and infrastructure investment in the telecommunication sector. The empirical analysis we conduct is based on a comprehensive data set covering 180 fixed-line and mobile operators in 25 European countries over 10 years and employs a newly created indicator measuring regulatory intensity in the European countries. We carefully treat the endogeneity problem of regulation by applying instrumental variables and find that tough entry regulation (e.g. unbundling) discourages infrastructure investment by entrants but has no effect on incumbents in fixed-line telecommunications. We do not find significant impact of entry regulation on investment in mobile telephony.

Keywords: Telecommunications, Access Regulation, Unbundling, Investment, European Union

JEL Codes: C51, L59, L96

* ESMT European School of Management and Technology, Schlossplatz 1, 10178 Berlin, Germany, tel.: +49 (0)30 21231-1047, fax: +49 (0)30 21231-1281, e-mail: grajek@esmt.org.
Financial support from the Deutsche Telekom AG is gratefully acknowledged. The opinions expressed are exclusively those of the authors.

1. Introduction

The regulatory framework for eCommunications ("the regulatory framework") defines the fundamentals of competition for the European telecommunication sector and is currently under review by the European Commission. The issues addressed by the framework can be separated into two broad groups. First, the framework defines which market segments of the telecommunication sector should be put under an ex ante approach of regulation and which market segments should be left to ex post regulation, i.e. competition policy. This is the question of what is the optimal instrument - ex ante regulation or competition policy. Second, it defines and harmonises the rules for ex ante regulation between the European member states. This is the question of how to optimise the instrument of ex ante regulation.

The answer to both questions — what is the optimal policy instrument and how to optimise the instrument — is by and large determined by the trade-off between static and dynamic efficiency: low prices in the short term, enforced through access or price regulation or through effective competition, may support static efficiency but may hamper investment in infrastructure and new products in the long term, that is dynamic efficiency. A robust understanding of the trade-off between static and dynamic efficiency is therefore central to the review of the regulatory framework.

Interestingly, having more than 20 years of experience with regulating telecoms worldwide, policy makers, practitioners and scholars still do not agree on the ideal approach that would yield a right balance between static and dynamic efficiency. For instance unbundling, the leading regulatory solution both in Europe and the U.S. in the late 1990's, which consists of ensuring new entrants' access to the incumbent fixed-line infrastructure at the wholesale level, has been phased out in the U.S., while it is still dominant in Europe (Renda, 2007).

This study intends to add to the debate on dynamic — or long-term — effects of the regulatory framework a more careful assessment of the resulting infrastructure investments in the industry. For that purpose an extensive literature review of the debate is provided and an empirical framework, which allows for robust inference given available data, is put forward.

The literature overview in the next section starts with a general assessment of the link between competition and investment and continues with a discussion of the telecom sector's specificities. We discuss the trade-offs between static and dynamic efficiency of competition when network infrastructure is difficult or impossible to duplicate and whether retail competition can lead to facilities-based competition. We also review the incentives of

incumbents and entrants to invest in infrastructure and the way mandated access at a regulated price influences these incentives. Finally, we report empirical evidence on those issues.

Section 3 reviews empirical models of telecommunication's infrastructure investment in the economic literature and proposes an econometric framework for our analysis. The most important elements of this framework are:

- Structurally modelled dynamics of the investment process, which allow us to derive short-term and long-term effects of regulation.
- A careful treatment of the endogeneity problem of regulation with instrumental variables technique.
- Disaggregated level of analysis, which accounts for the fact that regulation is segment-specific; moreover, it allows for differential effects of regulation on the fixed-line and mobile segments, as well as on the incumbents and entrants.

This empirical framework puts relatively high requirements on data. Section 4 reviews existing data that will facilitate our empirical analysis. We concentrate on different measures of investment and regulation and highlight their advantages and disadvantages. We also provide a number of control variables for the investment analysis, as well as possible instrumental variables for the regulatory measures to address the endogeneity concerns.

Section 5 provides a non-technical discussion of our econometric results along with a simulated effect of access regulation on investment in the industry.

Section 6 concludes by summarising the debate on regulation and investment in the literature and discussing implications of our empirical results.

Finally, robustness checks, a more technical discussion of both the theoretical and the empirical model, as well as detailed description of the data used for the analysis is placed in the annex.

2. Previous Literature

The ultimate reason for regulating the telecom markets is to introduce competition, which is widely believed to enhance efficiency and thereby social welfare. In the static sense, competition reduces the market power of producers, which leads to lower prices and higher surplus for customers. Competition also disciplines producers in their use of resources thereby promoting efficient use of inputs and minimising waste. To gain a more complete picture of

the relationship between competition and welfare one needs, however, to extend the textbook analysis of static efficiency by dynamic considerations, in which innovation and investments are key.

2.1 The General Trade-Off in Competition on Investment

Simple models of competition suggest a negative relationship between competition and investment and innovation.[1] Models of product differentiation and monopolistic competition deliver the prediction that more competition — for instance through lower transportation cost or higher substitutability between the products — reduces post entry (or post investment) rents and thereby reduces the incentives of firms to enter a market or to invest in new products or better processes. This effect, which is called the Schumpeterian effect in the literature, is also the key driver of the relationship between competition and innovation in traditional models of growth.

Recent research indicates that the relationship is in fact more complex and can be characterised by an inverted U-relationship. At a relatively low pre-existing level of competition an increase in competition will foster investment and innovation. After a certain saturation point, however, further increases in competition will result in reduced investment levels. While the latter can be explained by the Schumpeterian effect described before, the positive effect is due to the incentive of the incumbent to escape competition by innovation: increased competition reduces a firm's pre-innovation rents by more than it reduces its post-innovation rents. In other words a firm can escape lower rents by innovation. Accordingly this effect is called escape effect.[2]

The combination of these two effects, the Schumpeterian effect and the escape effect, allows for a vast array of industry specific results, depending on the ex ante level of competition and the distance of the incumbent firms from the technology frontier. Complementarities between the various instruments of an effective national investment/ innovation system add complexity to this relationship.[3]

[1] Innovation can be interpreted as a specific form of investment, resulting in new or better quality products and services or in more cost efficient processes. But there are innovation specific issues, like information spillovers or the public good character of innovations which have to be taken into account. For the purpose of this overview we will abstract from those specificities and use the two notions interchangeably.

[2] See Aghion et al. (2005) and Griffith et al. (2006) for a survey of the literature. The *escape effect* is closely linked to the discussion of whether an incumbent or a potential entrant has higher incentives to innovate. See for instance Gilbert, R. and D. Newbery (1982).

[3] See Mohnen and Röller (2005) for an empirical analysis of these complementarities.

2.2 Facilities-based vs. Service-based Competition

In the context of telecommunications industries, the potential efficiency gains from competition can be severely hampered by parts of the infrastructure that have natural monopoly properties. The local loops, which connect individual households to the local switch, are the most often cited example of such infrastructure. Duplication of the copper wires constituting the local loops is prohibitively expensive, at least for the purpose of an alternative supply of traditional telecommunication service. Both in Europe and the U.S. a typical solution to this infrastructure bottleneck was the introduction of a mandated access to the incumbent telephone network by means of unbundling and sharing of the local loop.[4] The mandated access facilitates the so-called service-based competition, in which the entrant is able to compete with the incumbent in the retail market by leasing the local loop at some regulated price. This is very different from facilities-based competition, in which both the incumbent and the entrant own the essential infrastructure and no leasing arrangements are required.[5] Most of the commentators are persuaded of the advantages of the facilities-based competition in terms of variety, keen prices and innovation, whereas the service-based competition seems to provide no other benefits than keen prices through the regulator-promoted access (Cave, 2004). Empirical evidence from the broadband market in Europe indeed suggests that in particular infrastructure competition between DSL and cable TV providers had a significant positive impact on the broadband deployment (Höffler, 2007).

In the context of the present study, it is very important to distinguish between the facilities-based and the service-based competition due to their potentially very different impact on innovation and investments. In particular, pooling the fixed-line and mobile infrastructure investment might give an inaccurate picture of the response of investments to the regulator-promoted competition, as mobile telephony, in contrast to fixed-line telephony, is characterised by full-fledged facilities-based competition.

[4] See Renda (2007) for a recent overview of the industry and the regulatory trends on both sides of the Atlantic.

[5] Although leasing of infrastructure is no longer required, interconnection of the competing networks and bilateral access prices remain an issue. Regulatory concerns under this two-way network are, however, significantly reduced as compared to one-way networks, when the entrant must seek access to the incumbent's essential facilities (Valetti, 2003).

2.3 Static and Dynamic Efficiency of Access Regulation

If access regulation reduces the monopoly power over the telecommunications infrastructure, then it also reduces the rent that can be earned on an investment in this infrastructure. Access regulation based on a simple cost recovery rule, while encouraging efficient utilisation of infrastructure, risks discouraging investment (Valetti, 2003). Therefore, there seems to be a trade-off between optimal regulation in a static and in a dynamic sense.

The increased static efficiency due to access regulation seems to be undisputed.[6] There are, however, conflicting views and research results on the impact of access regulation on investments in telecommunications, although the majority of the scholars tends to agree that access regulations in fact undermines infrastructure investment. This view is also reflected in a recent shift of the Federal Communications Commission (FCC) in the U.S. away from the access regulation (Renda, 2007).

In the context of dynamic efficiency, there is no firm theoretical argument in favour of access regulation. In the game-theoretic models of Foros (2004) and Kotakorpi (2006), service-based competition may encourage investment by the incumbent if it brings more variety and innovative services thereby boosting end-consumers' demand. Some scholars argue along these lines to conclude that lower access prices actually increase investment in facilities (Hassett et al. 2003; Willig, 2003). It is crucial though that profit from this increased market could be appropriated by the incumbent through high enough (possibly unregulated) access charges. This explains why Wallsten (2005) finds that Unbundled Network Element (UNE) regulations are negatively correlated with broadband deployment in the U.S., but resold lines are positively correlated with it.[7] The cost-based access charges promoted by the U.S. and the European regulators have been criticised, however, for being too low (e.g. Pindyck, 2004).[8]

Nevertheless, there exists some empirical support of access regulation promoting infrastructure investment in both U.S. and Europe. After analysing the sample of 41 Incumbent Local Exchange Carriers (ILECs) over 1994-1998, Chang et al. (2003) reports that the percentage of digital lines is negatively correlated with the access price and concludes that low access prices spur incumbents' investments. Similarly, the London Economics (2006) study for Europe finds that telecoms investments are higher when regulatory regimes perform

[6] Hausman and Sidak (2005) report, however, that mandatory unbundling resulting from the Telecommunications Act of 1996 does not appear to have decreased retail prices of the U.S. telecommunications services.

[7] UNE regulated prices were supposed to reflect the cost an incumbent would incur to provide each network element, while resale prices were supposed to be a discount from retail prices reflecting the incumbent's avoided costs of providing certain customer services. Hence, it was generally less expensive for competitors to provide service through UNE lines.

[8] See Valetti (2003) and Vogelsang (2003) for a general overview of the access pricing and its possible effect on innovation and investment.

better.[9] Li and Xu (2004) also find a positive effect of competition on telecommunications investments in a study based on a panel of 177 countries. There are two main drawbacks of these studies, though: i) Correlation is often taken as evidence for causation ignoring endogeneity concerns;[10] and ii) Data for the analysis — both regulation and investment measures — is often very aggregated, which ignores specificities of the fixed-line and mobile sectors, as mentioned earlier. These drawbacks cast severe doubts on robustness of these empirical findings. To the best of our knowledge, there are no empirical studies of investment and regulation in the telecom sector that address both these drawbacks at the same time.

On the other hand, the arguments that mandated access coupled with cost-based access charges undermine innovation have a relatively strong theoretical underpinning and include: i) Lowering the option value of the incumbent's investment, ii) Shifting the burden of risk from the entrant to the incumbent and iii) Increasing the incumbent's cost of capital. The first argument raised by many scholars (e.g. Haring and Rohlfs, 2002; Pindyck, 2004) says that by limiting future streams of profits on an investment access regulation decreases the Net Present Value (NPV) of the investment and thereby makes it less attractive for the investor.[11] In fact, this intuition drives the result that a lower unbundling price reduces the incumbent's and entrant's infrastructure investment in many formal models (e.g. Foros, 2004; Zarakas et al. 2005; Kotakorpi, 2006; Vareda, 2007).

The second argument points to the fact that the telecommunications infrastructure investment is highly uncertain and that the cost-based access charges do not take full account of that (e.g. Jorde et al., 2000; Haring and Rohlfs, 2002; Valetti, 2003; Pindyck, 2004; Baake et al., 2005). Instead, the mandated access charges give a risk-free option for entrants to lease the infrastructure and exploit the regulatory arbitrage between wholesale and retail prices. This in fact adversely affects the ex ante incentives of entrants to invest in their own infrastructure.

Besides, by shifting the burden of risk from the entrant to the incumbent, the cost-base access regulation may also increase the incumbent's cost of capital (Jorde et al., 2000) by diminishing its ability to invest. In particular, entrants are more likely to lease the local loops in case of unfavourable realisation of the uncertainty, i.e. when demand for telecommunications services turns out to be weak. Alternatively, when the demand is strong, higher prices for the services will afford entrants to roll out their own networks. Because the cost-base access charges undercompensate the incumbent, its returns will suffer in times of

[9] The performance of the regulatory regimes is measured by the OECD regulatory index, which is composed of three sub-indices: i) legal barriers of entry, ii) level of public ownership in telecoms and iii) market shares of entrants.

[10] This issue is addressed in more detail in the section on determinants of regulatory outcomes.

[11] See Pindyck (2004) for an introduction to the concept of option value and its application to investments in telecommunications infrastructure.

recession and improve during an expansion. This increased volatility of incumbent's returns on assets relative to the market has to be compensated by higher returns on its stocks for the investors, which increase incumbent's cost of equity. In their econometric analysis based on U.S. data Ingraham and Sidak (2003) found empirical support for this hypothesis.

There also exists some more general empirical evidence of the discouraging impact of access regulation on the investments in telecommunications. After analysing the industry trends in the U.S., the U.K., New Zealand, Canada and Germany, Hausman and Sidak (2005) concluded that mandatory unbundling did not spur infrastructure investments neither by incumbents nor by entrants. Using data from the U.S. over the period 2000-2001, Crandall et al. (2004) estimated that the share of the entrants' lines that are facilities-based is lower in the U.S. where the local loop rental rates are lower. Applying similar methods with European data, Waverman et al. (2007) demonstrated a strong substitution from broadband offered over alternative access platforms toward unbundled-loop-based offerings when local loop prices were low. This suggests that unbundling decreases entrant's investment in infrastructure and as a consequence facilities-based competition is lessened. In the same way, Eisner and Lehman (2001) found that states with lower unbundling rates experienced less facilities-based entry. Other studies found also a detrimental effect of unbundling policies on incumbent's investments (Haring, Rettle, et al. 2002; Crandall and Singer, 2003). Finally, Wallsten (2006) estimated the impact of local loop unbundling on broadband deployment to be insignificant or even negative in the OECD countries. These econometric studies ignore, however, the endogeneity of regulatory policies, which may significantly bias the results.

2.4 Can Service-based Competition Lead to Facilities-based Competition?

Proponents of the access regulation stress that although low access fees may not promote infrastructure investments, they do allow the entrants to climb the first rung of an investment ladder (Cave and Vogelsang, 2003; Cave 2004). In the first step an entrant would be able to attract its installed base of subscribers and gain a better understanding of the demand and the costs by leasing the parts of the incumbent's infrastructure that are very costly to duplicate. After accomplishing this first step, an increase in access charges together with technological progress and falling costs should encourage the entrant to roll out its own network and start the facilities-based competition. This logic is consistent with the formal model of Bourreau and Dogan (2005), who show that the optimal access charge from incumbent's viewpoint

would be prohibitively high during the time when there is no effective threat of facilities-based entry due to high investments costs. This access charge would then decrease over time together with technological progress, which renders the entry less expensive. By following this strategy, the incumbent could delay the facilities-based entry and at the same time extract maximum rent from the entrant.

The “ladder of investment” approach has been, however, heavily criticised by some scholars for not being effective in practice. After analysing industry trends, Hausman and Sidak (2005) found no evidence in favour of the “ladder of investment” hypothesis in the U.S., the U.K., New Zealand, Canada, and Germany. Hazlett and Bazelon (2005) reached the same conclusion based again on the U.S. data. We are not aware, however, of any systematic econometric study that would support or reject this hypothesis.

3. Empirical Framework

In this section we review the empirical models of infrastructure investment and regulation that have been used in the literature. In particular, we highlight the theoretical underpinning, as well as the treatment of the endogeneity problem in these models, as this is fundamental for the proper interpretation of the results. Next, we present our preferred model to be used in the subsequent analysis.

3.1 Existing Empirical Models of Infrastructure Investments

Most of the existing empirical models on infrastructure investments take explicitly or implicitly a reduced-form approach, in which investments or infrastructure level depends on a set of supply and demand characteristics (e.g. Chang et al., 2003; Crandall et al., 2004; Henisz and Zelner, 2001; Höffler, 2007; Wallsten, 2003). The only exception that we are aware of is the model of Röller and Waverman (2001), who estimate both the supply of and the demand for telecommunications infrastructure. One advantage of their structural approach is that it allows a predicting impact of the variable of interest separately on the demand and the supply. This might be important for instance if one wants to test the specific hypothesis — introduced in the literature review section — that access regulation boosts the end-consumer demand for infrastructure via innovative services of the retail competitors. This boost of demand may in

turn induce more infrastructure investments by the incumbent.[12] In contrast, the reduced-form model would be able to deliver only an estimate of the aggregated effect of demand and supply on the equilibrium level of infrastructure.

Another dimension that differentiates the empirical models is the use of dynamics. Static models assume that all relationships in the model occur immediately in a given period of time. One could, however, easily imagine that some effects might be postponed in time or occur with a different strength in the short and in the long run. The most simple way to account for these dynamics is to introduce lagged explanatory and lagged dependent variables to the model (e.g. Alesina et al., 2005). Greenstein et al. (1995) put more structure into the hypothesised dynamic process by considering a long-term equilibrium relation along with an adjustment equation. By doing so they derived an infrastructure equation with structural lags. For an investment model it is very important to incorporate these dynamics. Some of the investment decisions can be taken immediately and will add to the observable short term effects. Some of these decisions need adjustment time and will therefore only gradually translate into real effects. Hence, the accumulated effect can significantly differ from the short term effect. A static model, which basically captures the short term effects, can significantly misrepresent the true relationship.

Endogeneity issues also proved important in the models with regulatory variables. There are two main sources of endogeneity: reverse causality and omitted variables. Crandall (2005, p.71) points out the reverse causality problem by showing that the U.S. access prices in 2002 are negatively correlated with 1996-99 capital spendings of incumbent telecoms companies. Running a regression of capital spending on access prices, it is then very likely to find that lower prices are correlated with higher capital spendings, which may have nothing to do with the true causal effect of regulation on investments.

Omitted variables might also lead to endogeneity and hence biased estimates. Considering for instance a hypothetical world, in which regulation has no effect on investment, but facilities-based competition has a significant positive effect. Being aware of these competition effects but ignorant about own powerlessness, the regulators might choose a “hand-off” approach when the facilities-based competition is strong. An empirical analysis of the effect of regulation on investment ignoring the competition would then find a negative effect of regulation on investment, when in fact there is no such effect. A careful choice of variables and panel data techniques help to mitigate the omitted variable problem. More generally, the endogeneity issues can also be tackled with the instrumental variables (IV) techniques.

[12] This boost of demand should not be mistaken with moving along the demand curve by forcing the prices to fall. It should rather be understood as an outward shift of the demand curve.

Most of the above-cited studies acknowledge the potential endogeneity of regulation without controlling it. Studies that address the endogeneity problem by using IV-techniques include Gual and Trillas (2004), Gutierrez (2003) and Li and Xu (2004). Small sample size and high aggregation of the data, however, undermine robustness of the results in these studies.

3.2 Determinants of Regulatory Outcomes

All regulatory outcomes including unbundling policies and mandated access prices are the effect of political and administrative processes, which can interact with the investment decisions by firms. This is crucial for the econometric modelling of the investments and known in the econometric literature as endogeneity problem. Ignoring the endogeneity might lead to severe biases in the empirical results and difficulties for interpretation of the results, as highlighted in the previous section. In order to account for the endogeneity it is important to know what the determinants of regulatory outcomes are.[13]

While the relevance of this argument was pointed out already by Stigler (1971), only recently empirical studies established a close link between political and institutional factors and the design and the effectiveness of regulation. For instance, Neven and Röller (2000), Duso and Röller (2003) and Duso (2005) show that political and institutional factors explain a substantial part of the variation in subsidy levels between various EU countries, the degree of deregulation achieved in various OECD countries in the mobile telecommunication industry and price regulation in the U.S. mobile industry, respectively. These political and institutional factors include governments' general ideologies (left vs. right wing), governments' attitudes toward market regulations, electoral systems, political systems (presidential vs. parliamentary), accountabilities and independence of the regulatory agencies, as well as electoral campaigns' contributions. While the list of scholars, who seek to explain the regulatory policies, is much longer than the one cited here, the list of explanatory variables used typically includes the above variables.

As also shown in the above cited studies, one additional factor which explains the regulatory policies is the performance of the regulated market itself. In fact, this is one potential source of endogeneity in models that empirically estimate the relationship between the performance — measured for instance by investments — and the regulatory measures. If the causality does not only go from regulation to performance, but also in the reverse direction, then the simple

[13] If we find the regulatory determinants that are not correlated with the dependent variable – infrastructure investment in this study – we can use them as instrumental variables.

correlation between these two variables will reflect an average between these two causal relationships. For instance Crandall (2005, p.71) shows that the U.S. access prices in 2002 were negatively correlated with 1996-99 capital spending of incumbent telecoms companies, which suggests that regulators exploit investment ex post by reducing the rate at which the investing company is obliged to lease its network to competitors.

3.3 Determinants of Infrastructure Investments

Based on the literature reviewed in previous sections we identify four groups of variables that are likely to affect the infrastructure investment of a firm: i) demand shifters, ii) cost shifters, iii) competitive pressure and iv) regulation. The first group consists of variables affecting consumer demand for telecommunications infrastructure. These variables include consumer wealth typically measured by GDP per capita.

The second group covers investment cost shifters. Because the density of households determines to a large extent the costs of building the local loops, a natural cost measure is the population density and the level of urbanisation. The costs of labour and capital obviously play an important role as well. The cost of labour in the construction sector seems particularly relevant for the infrastructure investment and the debt level of a firm may serve as a good proxy of its cost of capital. Many commentators also point to the dot.com bubble, which burst in 2001, severely affecting the investments that the telecoms operators could afford. The stock market bubble can be accounted for by means of time period (year) dummy variables.

The third group of variables comprises measures of competitive pressure.[14] In particular, investment incentives of telecom companies can be influenced by facilities-based competition from alternative platforms. One such measure used in the literature is cable TV penetration, as cable broadband offerings directly compete with DSL broadband access over fixed-lines. By the same token, the number of main lines in a country constitutes a measure of competitive pressure in mobile telecoms.[15]

Regulatory policies constitute the forth group of relevant variables. Among them entry regulation including unbundling and sharing of the local loop are most heavily disputed.

[14] A sustainable competition is an ultimate goal of the telecom sector's regulation in Europe, but the two should not be confused.

[15] It is important to stress here that the optimal choice of explanatory variables should not aim to explain as much variation in the investment variable as possible, but rather minimise omitted variable problems thereby contributing to the accuracy of estimates on the regulatory variables. Inclusion of variables that might be correlated with investment levels as well as regulatory policies (like the installed cable TV infrastructure) is then crucial.

Finally, but most importantly, we have identified a set of instrumental variables in order to control the endogeneity of regulatory policies. The following variables have been identified as potential instruments in our estimations:

- Political variables: Political ideology of the government, attitude of the government toward European integration, attitude of the government toward regulation, as well as the level of checks and balances constraining the discretion of politicians' and bureaucrats' decisions.
- Neighbouring markets: We also consider using the level of regulation in other European countries as possible instrument.

3.4 The Econometric Model

The econometric model that we propose follows Greenstein et al. (1995). It is a partial adjustment model, in which the current infrastructure stock is a weighted average of the long-run desired stock and of the lagged stock value, where the weights reflect the speed of adjustment to long-run equilibrium.

As shown in the annex, the partial adjustment model yields the following estimation equation:

$$Infr_{kjt} = \alpha_0 + \alpha_1 Infr_{kjt-1} + Demand_{kjt}\boldsymbol{\beta_1} + Cost_{kjt}\boldsymbol{\beta_2} + Comp_{kjt}\boldsymbol{\beta_3} + Reg_{kjt}\boldsymbol{\beta_4} + v_{kjt}. \qquad (1)$$

$Infr_{kjt}$ reflects the stock of infrastructure for firm j in country k in time period t and $Infr_{kjt-1}$ is the stock of infrastructure in the previous period. $Demand_{kjt}$, $Cost_{kjt}$, $Comp_{kjt}$ and Reg_{kjt} stand for the four groups of variables that determine the infrastructure investment, as identified in the previous section, and $\boldsymbol{\beta_1}$ through $\boldsymbol{\beta_4}$ denote the respective four groups of coefficients for these variables. Finally, v_{kjt} is a usual error term, which captures the variation in the infrastructure that is not explained by the model.

The lagged dependent variable $Infr_{kjt-1}$ in equation (1) distinguishes this model from standard static linear regression models. The inclusion of the lagged dependent variable follows from the assumption that firms do not immediately adjust the level of infrastructure to changing market conditions. Instead, the adjustment is distributed over years and in each year only a fraction of an optimal long-term investment is actually undertaken. The investment process is then assumed to exhibit certain inertia, which is reflected by the coefficient α_1.

The estimation of eq. (1) provides then information on two aspects of the investment process: First, the estimates of $\boldsymbol{\beta_1}$ through $\boldsymbol{\beta_4}$ provide the short-run effects of regulatory and economic variables on the stock of infrastructure; second, α_1 reflects the speed of adjustment and, as a consequence, the long-run effects on infrastructure.

4. Data

This section presents the data we use in the analysis. First, we discuss in some detail the main variables of the study, i.e. the investment and the regulatory variables. Second, we present the set of explanatory variables and the variables that we use as instruments to address the endogeneity of regulation.

4.1 Investment and Regulation Measures

In our search for the investment variable that would facilitate our empirical analysis we followed three main criteria: i) proximity to the real infrastructure or infrastructure investment, ii) extensive coverage in terms of European countries and time periods, iii) sufficient level of disaggregation in terms of geographical markets and service segments (fixed-line vs. mobile).

The stock of infrastructure and the level of infrastructure investment can be measured in many ways. One variable often used in the literature is the physical amount of infrastructure measured in the number of main telephone lines, kilometres of fibre optic cables, share of digital lines, etc. The number of main telephone lines is a readily available variable for European markets, but the more detailed measures are not.

We therefore concentrate on financial measures of infrastructure stock investments, which include tangible fixed assets, Property, Plant & Equipment (PPE), additions to (tangible) fixed assets, additions to PPE, capital expenditures (CAPEX), and country-level aggregated investments in telecoms. These variables differ in terms of the proximity to real infrastructure investments, as well as their availability, sample coverage and the level of aggregation, which creates trade-offs. In short, one could either choose a more precise variable or a variable with larger sample coverage. For instance, firm-level CAPEX from the Osiris database is a very accurate measure of infrastructure investment and is disaggregated to the country level, but

only some 60 data points are available. The ITU database in turn offers country-level investment figures separately for fixed-line and mobile telecoms, but only 180 data points are available over the period 1990-2006. Moreover, ITU figures do not distinguish between the incumbents' and the entrants' investments. On the other hand, there are over 1.000 observations for tangible fixed assets available from Amadeus database (supplemented with some figures from Osiris). These tangible fixed asset figures are geographically well-defined (country-level) and come from more than 200 firms (incumbents and entrants; fixed-line and mobile sector) in more than 30 European countries over the period 1997-2006, which offers a very rich variation both across countries and in time. The disadvantage of tangible fixed assets as a basis for investment measures is that they are affected by revaluations, as well as mergers and acquisitions (M&A), which do not count towards real infrastructure investments. This will not be a problem for an econometric model if the M&A and the revaluations are not correlated with the explanatory variables of the model. They will then merely enter the error term vkjt. To the extent that there are some spillover effects between merger control and regulatory policy in a country, however, there will be an endogeneity bias in the estimated coefficients on the regulatory variables. This problem can be addressed by including a variable reflecting M&A activity of the firm.[16]

The current infrastructure stock $Infr_{kjt}$ in equation (1) is then measured by a firm's tangible fixed assets deflated by the Producer Price Index (PPI) for telecom equipment. This measure fits well with the econometric framework of our analysis and allows us to gain detailed insights into the investment process in the industry. It was taken great care to assure that our infrastructure measure corresponds well to the geographic markets, which are defined by countries' borders, as well as to the market segments (mobile vs. fixed-line). The list of companies in our sample together with a detailed description of how the infrastructure measure was constructed is reported in the annex.

The regulation variables that we use in our analysis come from Plaut Economics (Zehnhäusern et al., 2007). The advantage of Plaut's regulatory index is its detailed information on different regulatory measures in the telecom sector and the comprehensive coverage in terms of countries and years. It is available for all 27 EU countries over the period 1997-2006. The index is divided into sub-indices, for price regulation, quantity regulation, market entry regulation and for miscellaneous other regulations. Price regulation scores the interconnection-regime and existence of sector-specific retail price-regulation with regards to

[16] Since political and institutional variables may affect the merger control as well as the regulation, the IV estimator based on these instruments, which we apply as a general remedy to endogeneity, may not be immune to this particular endogeneity problem.

fixed and mobile telecom, as well as potential F2M-termination regulation. Quantity regulation scores the existence of a Universal Service Obligation burden for incumbents or other telecom companies by the NRA and the existence of meet-demand-clauses for specific products or services at regulated (retail) prices or regulatory requirements regarding the coverage of mobile telephony-services. Market entry regulation scores the existence of regulated vertical separation or an accounting separation obligation, existence of various types of regulated access to the incumbent's network and the number of network-based 2G/3G mobile licenses. Finally, miscellaneous regulation scores the percentage of government-ownership of the incumbent, existence of a "golden share", access regulation asymmetry between DSL and cable network providers, existence of a sector-specific regulation in favour of protecting the environment, etc.

In line with the current political and scholarly debates, we focus on market entry regulation among all regulatory tools used in the telecom sector. The modular construction of the Plaut's regulatory index allows us to construct segment-specific indices for the mobile and fixed-line segments. Our regulatory index for the fixed-line segment is an average of indicators referring to the existence of regulated vertical separation and an accounting separation obligation, as well as the existence of regulation regarding the full unbundling, line sharing, bitstream access and subloop unbundling of the fixed-line incumbent's local loop.[17]

For the mobile segment, our regulatory index is an average of indicators referring to the number of network-based 2G and 3G mobile licenses and the constraints on the trade of frequencies.[18] Besides the lower number of sub-indices available, the indicators for mobile telephony have to be treated with some caution for the purpose of this study. The number of network-based licences focuses – in contrast to the indicators used for fixed-line telephony - on facility-based entry, for which economic theory predicts significantly different results. But alternative indicators addressing non-facility based entry regulation, like the existence of mobile virtual network operators for instance, are not available. The interpretation of the results has to take this into account when comparing the outcome for fixed-line telephony and mobile telephony.[19]

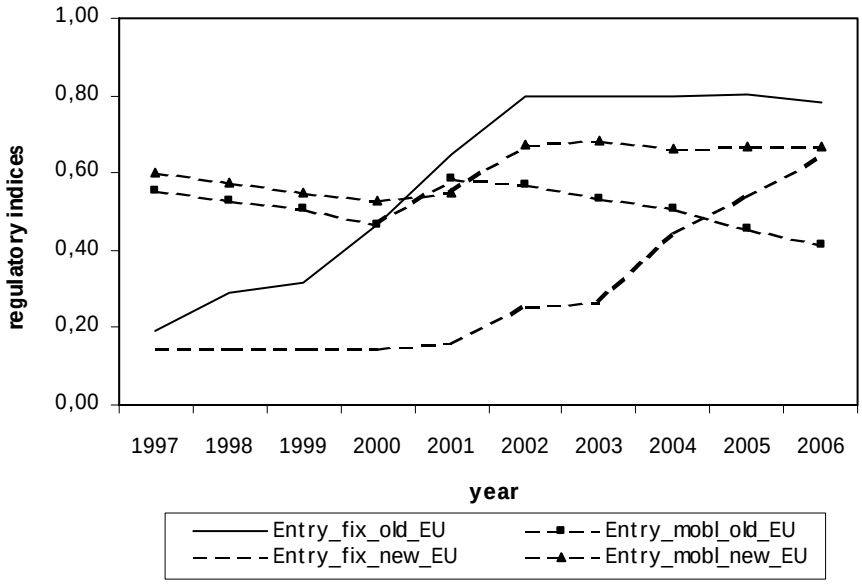
[17] The indicators entering our regulatory index for the fixed-line segment correspond to the keys 11 through 16 and 22 of the Plaut's index.

[18] The indicators entering our regulatory index for the mobile segment correspond to the keys 17 through 19 of the Plaut's index.

[19] A further limitation of the number of licenses granted as an indicator for entry regulation is that it does not include information on coverage obligations linked to those licenses. The impact of licenses which oblige the licensee to a specific level of investment will have a significantly different effect on investment levels than licensees granted without such a coverage obligation.

Figure 1 shows the evolution of the European telecom sector's entry regulation, as defined in our analysis, over the last 10 years. The "old" EU members (EU 15) experienced growing regulatory intensity in the fixed-line segment, which levelled-off in 2002. The new member states, in contrast, did not introduce any substantial measure promoting entrants to the fixed-line telephony until the eve of the 2004 EU accession.

Figure 1: Entry Regulation in Fixed-line vs. Mobile Telephony in EU Markets



Source: Authors' calculations based on data from Plaut Economics

The regulation of mobile telephony, mainly driven by licensing, was much more stable over time and equal across the new and the old member states. The fall in the index for the old Europe starting in 2001 can be attributed to the new 3G mobile licenses being granted as the new technology made its inroad to the markets.[20]

To sum up, the main variables of our study — stock of infrastructure and entry regulation in the mobile sector — are sufficiently disaggregated to pinpoint the differential impact of regulation on investments across the industry's segments, as suggested by the theoretical literature. Having such a rich firm-level data will also allow us to study the asymmetries between incumbents and entrants. Finally, the large coverage of our sample in terms of geographical markets and years facilitates a robust econometric analysis.

[20] More licenses are attributed to less regulation (more competition) by the index.

4.2 Other Control and Instrumental Variables

The definitions and sources of all variables used in the estimation of equation (1) are reported in table1. Table 2 reports the descriptive statistics. The first group of explanatory variables, referred to as main controls in the tables, includes demand shifter (GDPpc) along with an array of variables controlling for different types of companies in our sample. In particular, we distinguish between mobile operators from fixed-line operators and incumbents from entrants among fixed-line operators.[21] Because we could not obtain data for domestic fixed-line infrastructure for 10 out of 25 fixed-line incumbents in our sample, the infrastructure measure includes other operations of these companies as well, most importantly their mobile telephone operations. The Multisec indicator variable accounts for this.

Given the measure of infrastructure that we apply, it is important to control for M&A activities of the companies, as mentioned in the previous section. M&A transaction data was obtained from the SDC Platinum M&A database. Updated daily, the database offers detailed information on merger transactions including acquirer and target profiles, deal terms, financial and legal advisor assignments, deal value and deal status. This database includes alliances with a deal value of more than one million USD, thus ensuring that the overwhelming majority of mergers are covered. Mergers which took place in the telecommunications services industry in the EU region were selected and matched to our firm-level data set. Hereby, care has been taken to identify geographical ties of the transactions performed by multinational companies. Our final sample of merger transactions contains information on 229 completed deals announced during the period from 1997 to 2006 which were carried out by 54 firms. The values of the merger transactions were determined, while for multiple transactions by the same company in a given year, the sum of deal values has been computed correspondingly.

[21] We ignore the distinction between incumbents and entrants in the mobile telephony, because the asymmetries between them are far less important in practice. In particular, mobile entrants are not granted one-way access to the incumbents' network.

Table 1: Definition of Variables

Variable	Definition	Source
Dependent variable:		
Infr	Tangible fixed assets in domestic sub-sector (mio €, 2000 prices)	Amadeus Osiris
Main controls		
Mobile	Dummy = 1 if company is a mobile phone operator	Amadeus Osiris
Incumb	Dummy = 1 if company is an incumbent PTE (fixed-line)	Amadeus Osiris
Entrant	Dummy = 1 if company is a fixed-line entrant	Amadeus Osiris
Multisec	Dummy = 1 if assets of incumbent PTE include those employed in other than fixed-line operations (most importantly – mobile telecommunications)	Amadeus Osiris
M&A	Value of M&A transactions (mio €, 2000 prices)	SDC Platinum M&A
GDPpc	Gross domestic product per capita (€, 2000 prices)	World Bank's WDI
Regulation:		
EntryFix	Index of entry regulation in fixed-line markets	Plaut Economics
EntryMob	Index of entry regulation in fixed-line markets	Plaut Economics
Cost shifters:		
Labour	Annual index of labour input cost in construction	Eurostat
Debt	Ratio of long-term debt to total assets	Amadeus Osiris
PopDens	Pop dens Population density (people per sq. km)	World Bank's WDI
Competition:		
CompFix	Penetration rate of cable TV (households passed by cable)	OECD Communication Outlook
CompMob	Main telephone lines (fixed-lines) per 100 inhabitants	ITU World Telecom/ICT Indicators
Instruments:		
EntryFixNeighbour	Index of entry regulation in neighbouring fixed-line markets defined as average regulation in corresponding European countries	Plaut Economics
EntryMobNeighbour	Index of entry regulation in neighbouring mobile markets defined as average regulation in corresponding European countries	Plaut Economics
Regul	Measure of government's attitude toward market regulation	Manifesto Project
Rile	Right-left position of government	Manifesto Project
Europ	Measure of government's attitude towards European integration	Manifesto Project

Table 2: Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Dependent variable:					
Infr (mio €, 2000 prices)	1083	762.27	1,913.70	.0037	22,896.69
only Mobile	457	824.74	1,375.40	.0222	10,771.44
only Incumb	141	2,483.77	4,118.25	.0545	22,896.69
only Entrant	485	202.95	564.40	.0037	5,985.37
Main controls:					
Mobile	1083	.42	.49	0	1
Incumb	1083	.14	.35	0	1
Entrant	1083	.44	.49	0	1
Multisec	1083	.07	.26	0	1
M&A (mio €, 2000 prices)	1083	194.73	2,177.04	0	44,883.18
GDPpc (€, 2000 prices)	1083	16,016.05	8,950.33	1,450.22	43,357.70
Regulation:					
EntryFix	1083	.5458	.2837	.1428	.8571
EntryMob	1083	.5458	.1622	.1666	.8666
Cost shifters:					
Labour	1071	106.3	9.7	65.5	168.4
Debt	959	.22	.51	0	5.09
PopDens	1083	34.79	27.46	10.31	99.32
Competition:					
CompFix	702	52.52	28.35	0	100
CompMob	1027	.4673	.1331	.1505	.7576
Instruments:					
EntryFixNeighbour	1083	.5168	.2497	.1428	.7802
EntryMobNeighbour	1083	.5100	.0705	.4230	.7333
Regul	935	1.64	1.01	0	4.47
Rile	935	1.72	8.23	-12.64	28.47
Europ	935	2.68	1.73	-.78	7.18

The other control variables used in our analysis include various cost shifters and competition measures. Population density and labour costs in construction reflect the costs of infrastructure deployment. Furthermore, the debt ratio of a company may affect the financial conditions under which the infrastructure investment is financed. In short, the cost of capital may increase with the debt ratio leading to less investment. Finally, our competition measures are defined as penetration rate of cable TV and main telephone lines per capita for the fixed-line and mobile telephony, respectively.[22] Table 1 and table 2 report also the instrumental variables that we used to account for endogeneity of the regulation. First, we construct two geographical variables, EntryFixNeighbour and EntryMobNeighbour, capturing the average level of entry regulation in neighbouring markets. Moreover, in defining the neighboring

[22] Missing values in the time series for penetration rate of cable TV were filled by linear interpolation; 25% out of the 702 observations in the cable TV series were constructed this way.

markets we distinguish between the "old" EU (EU 15) and the "new" EU members, because the regulation of telecom sectors crucially depends on the EU accession, as illustrated in Figure 1. The neighbouring markets for Germany and Poland for instance are all other old EU members and all other new EU members, respectively.

Besides the geographical instruments, we also utilise variables measuring political environment in the European countries. The variables come from the Manifesto Project, which deals with different aspects of structures and performances of parliamentary democracies. The project focuses on quantitative content analyses of party manifestos from 50 countries covering all free democratic elections since 1945 to measure political positions of all relevant parliamentary parties.[23] The variables that we extract from this rich database are the overall policy positions of the government in terms of right versus left scale (Rile) and favouring market regulation (Regul), as well as the government's attitude towards European integration (Europ). The position of government is defined as the weighted average score of parties in the government and the weights are constructed as the proportion of parliamentary seats held by each party. In the election years, the government position is taken as the average position of the two consecutive governments weighted by the number of months in the office.

5. Empirical Results

This section contains a non-technical discussion of our results including a simulated impact of entry regulation on investment. A more technical discussion of the results, statistical properties of the estimated model and various robustness checks that we performed are presented in the annex.

Table shows the estimation results of the preferred specification of our econometric model. The continuous variables in the model are in logarithms, which allows us to interpret the respective coefficients as elasticities. The list of explanatory variables excludes the cost shifters and competition measures, as they turned out insignificant in the regressions. The results in table 3 are, however, robust to inclusion of these additional controls, as demonstrated in the annex. Country and year dummy variables are also included in the estimation, but are not shown in table 3 for brevity's sake. Country dummy variables capture all country-specific determinants of firms' investments, like consumer tastes, institutional environments, geographic characteristics, etc. to the extent that these do not change over time.

[23] See Klingeman et al. (2006)

The coefficients on the country dummy variables reflect then all these possible effects and are very useful as controls for possible omitted variables in the regression. Similarly, year dummy variables capture macroeconomic shocks that affect all firms in the analysis. For instance, the stock market bubble, which affected the investments that the telecom operators could afford, can be accounted for by year dummy variables.

The results in table 3 are obtained by Instrumental Variables (IV) estimation and show very high statistical significance. Very good statistical properties of the model and its robustness to alternative specifications are further documented in the annex. The dynamic specification of the model proved to be correct, as the coefficient on the lagged dependent variable is highly significant. The magnitude of the coefficient, which is very close to 1, means that the stock of infrastructure is highly time persistent. It also suggests that shocks to economic determinants of the stock of infrastructure have very persistent effects. A 10% increase in the stock of infrastructure due to a change in some economic conditions is followed by a further 9.4% increase in infrastructure in the next year, 8.8% in two years, 8.3% in three years and so on. The long-term effects are therefore much higher than the immediate effects according to our estimates.[24]

[24] Because the coefficient on lagged dependent variable is almost 1, we can actually redefine our dependent variable as log(Infr/Infr(-1)), i.e. the index of infrastructure, and interpret our results as infrastructure investment elasticities rather than infrastructure stock elasticities. The accuracy of this interpretation is higher when the time horizon is shorter.

Table 3: Dynamic Model of Investment: Instrumental Variables (IV) Estimation Results

Dependent variable: **Log(Infr)**

	(1)
Log(Infr) (-1)	0.94***
	(0.02)
Mobile	-0.63
	(0.49)
Incumb	-0.41***
	(0.08)
Multisec	0.27**
	(0.12)
Log(M&A) * I(M&A>0)	0.04**
	(0.02)
Log(GDPpc)	0.52**
	(0.26)
EntryFix[1] * Incumb	-0.02
	(0.21)
EntryFix[1] * Entrant	-0.44***
	(0.15)
EntryMob[1] * Mobile	0.87
	(0.82)
Observations	730
R-squared	0.96

Notes:
Robust standard errors are in parentheses.
The estimates for intercept, country-specific effects and year dummies are not shown.
* Significant at 10%; ** significant at 5%; *** significant at 1%.
[1] Endogenous variables: EntryFix and EntryMob; Instrumental variables: EntryFixNeighbour, EntryMobNeighbour, Regul, Rile, Europ and interactions thereof.

Our estimates for Mobile and Incumbent dummy variables further suggest that there is no significant difference in infrastructure investments between mobile phone operators and entrants into the fixed-line segment; however, relative to their infrastructure stock, the fixed-line incumbents' investments are on average 41% lower than the investments of entrants. This result is very intuitive, as the infrastructure stock of an average entrant in our sample is more than 10 times smaller (see table 2) implying that the relative dynamics are likely to be higher.

The controls for incumbents operating in multiple segments (Multisec) and M&A activities also turned out to be significant. The positive coefficient for Multisec is likely to be driven by the fact that the mobile telephone infrastructure shows higher dynamics than the fixed-line incumbents' infrastructure, as suggested by the estimates for Mobile and Incumbent dummy variables. The other interpretation is that the incumbents in the new member states, for which

the segment break down of infrastructure figures is typically not available, are “catching up” with the old member states’ standards. By checking the operators’ M&A activities we include only the observations with non-zero values, which is why the indicator variable I(M&A>0) is interacted with the M&A variable in table 3. The coefficient on this interaction variable is positive, as expected, but very low. This might reflect the fact that M&A deal values are largely driven by other than tangible assets.

The demand shifter measured by GDP per capita is also positive and significant in the regression, as expected. The estimated elasticity of 0.52 means that a 10% increase in average income per capita increases infrastructure investment by roughly 5%. Other control variables — cost shifters and competition measures — turned out insignificant in our regressions. One explanation for this is that we already control a large fraction of cost and competition differences between countries by means of country-specific effects. Therefore our additional explanatory variables do not seem to be precise enough to further explain the firm-level investment decisions.

Finally, turning to the regulatory variables — the focus of this study — we see a big asymmetry between segments as well as incumbents and entrants. Entry regulation seems to have no significant impact on investment in mobile telephony, but it significantly discourages investment in fixed-lines.

For the mobile segment, it has to be recalled that our regulatory index is an average of indicators referring to the number of network-based 2G and 3G mobile licenses and the constraints on the trade of frequencies. As pointed out before, the number of network-based licences focuses – in contrast to the indicators used for fixed-line telephony - on facility-based entry, for which economic theory predicts significantly different results. This has to be taken into account when comparing the outcome for fixed-line telephony and mobile telephony. A more consistent comparison of the estimation results for fixed-line telephony and mobile telephony would require an indicator focusing on serviced-based entry regulation in the mobile telephony sector. The existence of mobile virtual network operators on investment could be such an indicator, but is not available in the indicator set employed throughout this study.

In the fixed-line segment, entry regulation has a significant negative effect on the infrastructure investment by entrants. This result is consistent with theoretical predictions and existing empirical studies on regulation and investment in the telecom sector. In particular, it corroborates the finding in Waverman et al. (2007) that the intensity of access regulation in Europe negatively affects investment in alternative and new access infrastructure. It is

important to stress that Waverman et al. (2007) arrive at the same result as we do using a very different empirical approach. First, they measure access regulation as LLU prices rather than an indicator of existence of various types of access regulation and vertical separation of the incumbent operator. Second, they measure entrants' investment as the number of new broadband subscribers over alternatives like the LLU-based access platforms rather than a change in tangible fixed assets. Third, they utilise data aggregated to the country level rather than individual operators' data.

According to our estimate, an increase in the regulation index from 0 to 1 leads to a decrease in investment by 44%. To gain a better understanding of what this number means, we suppose that the NRA introduce an additional mode of regulated access to the incumbent's local loop; it could be for instance full unbundling, line sharing, or bitstream access. One such additional mode of access increases our regulation index by 0.14, leading to a decrease in investment by more than 6% on average.[25]

Another exercise we perform in order to quantify this effect is to simulate the aggregated loss in investment due to access regulation. The assumptions of the simulation are as follows:

- Countries: 25 EU members in our sample
- Time horizon: 5 years
- Access regulation: aggregated effect of all 4 means of access to incumbent's local loop (full unbundling, line sharing, bitstream access, subloop unbundling)

Our estimates suggest that the immediate effect of introduction of this access regulation — which corresponds to an increase in the regulatory index by 0.57 — is a lost investment in the amount of 25.1% of the entrants' infrastructure stock. In the following year the lost investments amounts to 23.6%, in two years – 22.2%, in three years – 20.9%, and in four years – 19.6%. The cumulative loss in investment over 5 years is then 111.5% of the entrants' infrastructure stock. In other words, our results suggest that the entrants would more than double their infrastructure over 5 years if they did not have regulated access to the incumbents' local loops. Accounting only for the companies in our sample — 80 entrants with an average infrastructure stock of €202.95 million — this loss amounts to € 4.1 billion in the first year and €18.1 billion over 5 years, which is equivalent to some 8.4% of the total telecommunication investment in Europe.[26]

In contrast to the entrants, the incumbents are not found to significantly decrease their investment as a result of entry regulation. One possible explanation of this finding is that

[25] Since the fixed-line regulation index consists of 7 indicators, each of the indicator accounts for about 0.14.

[26] To calculate this we took an average telecommunication investment per capita per year of € 100, which corresponds to some recent reports (OECD Communication Outlook 2007).

entrants are able to boost end customer demand due to increased variety and innovativeness of their information and communication services offered on incumbents' networks. In this case the lost profit margins of incumbents could be offset by the increase in total demand. It has to be highlighted though that the data used for the analysis does not cover investment in next generation access networks. To the extent that the investment in next generation networks is qualitatively different from upgrading the current infrastructure of incumbents, this result cannot be extrapolated to future investments.

6. Conclusion

This study adds to the debate on dynamic — or long-term — effects of the regulatory framework a more careful assessment of the resulting infrastructure investments in the industry. For that purpose an extensive literature review of the debate is provided and an empirical framework, which allows for robust inference given available data, is put forward.

The literature review reveals an important difference between facilities-based and service-based competition as goals for regulatory policies. Most of the commentators are persuaded of the advantages of the facilities-based competition in terms of variety, keen prices and innovation, whereas the service-based competition seems to provide no other benefits than keen prices through the regulator-promoted access. If facilities-based competition is an ultimate goal of proper regulation, then incentives to infrastructure investments become a key measure of success of this regulation. There are conflicting views and research results on the impact of access regulation — leading regulatory solution in the industry — on investments in telecommunications. The majority of the scholars tends to agree, however, that access regulation in fact undermines infrastructure investment, both by incumbents and entrants. This view is also reflected in a recent shift of the Federal Communications Commission (FCC) in the U.S. away from the access regulation.

The empirical analysis of infrastructure investment in telecommunications, which we conduct, is superior to existing studies in several dimensions:

First of all, the dynamics of the investment process are modelled structurally, allowing us to derive short-term and long-term effects of regulation. This approach also fits better to the available investment data (which are on infrastructure level) and allows the results to be linked to a macro-model of growth.

Second, a careful treatment of the endogeneity problem of regulation is proposed by identifying several potential instrumental variables. The following instruments are used for our estimation:

- Political variables: Political ideology of the government, attitude of the government toward European integration and attitude of the government toward regulation in general.
- Neighbouring markets: We also use levels of regulation in other European countries as possible instruments.

Third, we disaggregate the data so that different effects of regulation in mobile and fixed-line segments of telecommunications, as well as on incumbents and entrants, can be derived. For carrying out such an analysis disaggregated data of the regulatory indicator constructed by Plaut Economics is used along with detailed firm-level infrastructure measures.

Finally, our estimation is based on a comprehensive dataset covering 180 fixed-line and mobile operators in 25 European countries over 10 years. This allows a sample size of the overall regression of around 1000 observations.

Based on this methodology we derive the following main results:

First, estimating a static model (no lagged infrastructure stock variable is included) without controlling for the endogenity problem of regulation results in very different effects than what is found in a richer, statistically more appropriate approach. Using simplified approaches for policy advice can therefore be misleading.

Second, the dynamic specification of the model proves to be correct and robust. The magnitude of the coefficient on the lagged infrastructure variable, which is very close to 1, means that the stock of infrastructure is highly time persistent. It also suggests that shocks to economic determinants of the stock of infrastructure have very persistent effects. A 10% increase in the stock of infrastructure due to a change in some economic conditions is followed by a further 9.4% increase in infrastructure in the next year, 8.8% in two years, 8.3% in three years and so on. The long-term effects are therefore much higher than the immediate effects according to our estimates.

Third, we find that entry regulation discourages infrastructure investment by entrants in fixed-line telecommunications. According to a simulation based on operators in our sample, the introduction of a regulated access to incumbents' networks costs Europe a lost investment in the amount of 25.1% of the entrants' infrastructure stock in the first year. This loss accumulates over time and reaches 111.5%, which is equivalent to €18.1 billion, over 5 years.

In other words, our results suggest that the entrants would more than double their infrastructure over 5 years if they had no regulated access to the incumbents' local loops. In terms of the total telecommunication investment in Europe, the lost investment is equivalent to 8.4%, which is a significant amount.

Fourth, incumbents are not found to significantly change their investment as a result of entry regulation in fixed-line telecommunications. One possible explanation of this is that entrants are able to boost end customer demand due to increased variety and innovativeness of their information and communication services offered on incumbents' networks. In this case the lost profit margins of incumbents could be offset by the increase in total demand. It has to be highlighted that the data used for the analysis does not cover investment in next generation access networks. To the extent that the investment in next generation access networks is qualitatively different from upgrading the current infrastructure of incumbents, this result cannot be extrapolated to future investments.

Fifth, while entry regulation significantly discourages investment in fixed-lines by entrants, it seems to have no significant impact on investment in mobile telephony both by entrants and incumbents. This result may be due to the limited quality of the available indicator for entry regulation in mobile telephony, which comprises mainly the number of network-based licences. The number of network-based licences focuses – in contrast to the indicators used for fixed-line telephony - on facility-based entry, for which economic theory predicts significantly different results. But alternative indicators addressing non-facility based entry regulation, like the existence of mobile virtual network operators for instance, are not available.

Overall, the results of this study highlight the importance of using a robust empirical approach if econometric evidence is used for policy advice. Opposite to what is derived from simplified assessments we do not find any indications that entry regulation has a positive impact on investment. On the contrary and in line with the theoretical literature, in the fixed-line sector regulators are faced with an important trade-off, where we find a significant negative effect of entry regulation on the incentives of entrants to invest. Promoting market entry by means of regulated access might have the desired short-term effect of lower prices and more consumer surplus, but at the same time undermines the incentives of entrants to invest in their own infrastructure and thereby compromising on the long-term goal to establish facilities-based competition.

Appendix

A1. The Econometric Model: Derivation

The econometric model that we apply follows Greenstein et al. (1995). It is a partial adjustment model, in which the current infrastructure stock is a weighted average of the long-run desired stock and of the lagged stock value, where the weights reflect the speed of adjustment to long-run equilibrium.

In particular, we assume that $Infr^*_{kjt}$ reflects the long-run desired stock of infrastructure for firm j in country k in time period t. Let $Infr^*_{kjt}$ be given by

$$Infr^*_{kjt} = \boldsymbol{X}_{kjt}\boldsymbol{\beta}' + \varepsilon_{kjt}. \qquad \text{(A1)}$$

For brevity, $\boldsymbol{X}_{kjt}$ comprises all four groups of explanatory variables, as well as the constant term α_0. Current stock levels are given by the adjustment process:

$$Infr_{kjt} = Infr_{kjt-1} + \alpha_1'(Infr^*_{kjt} - Infr_{kjt-1}) + \mu_{kjt}. \qquad \text{(A2)}$$

Substituting eq. (A1) into eq. (A2), we obtain

$$Infr_{kjt} = \alpha_1 Infr_{kjt-1} + \boldsymbol{X}_{kjt}\boldsymbol{\beta} + v_{kjt}, \qquad \text{(A3)}$$

where $\alpha_1' = 1 - \alpha_1$, $\boldsymbol{\beta}' = \boldsymbol{\beta}/\alpha_1'$ and $v_{kjt} = \alpha_1'\varepsilon_{kjt} + \mu_{kjt}$.

Equation (A3) is identical with equation (1) in the main body of the text. Estimation of eq. (A3) provides information on two aspects of the investment process: First, the estimate of α_1' reflects the speed of adjustment. Second, the estimates of $\boldsymbol{\beta}'$ provide information on the effect of regulatory and economic variables on the long-run desired stock of infrastructure. The estimates

A2. The Econometric Model: Alternative Specifications

To check the robustness of our results we run additional IV regressions including a full set of explanatory variables. The results in table 4 show that they are generally robust to inclusion of the cost shifters and competition measures as additional explanatory variables. Because of missing observations, which tend not to show in these variables, however, our sample size drops from 730 to 445 observations and some statistical significance is lost.

Table 4 reports also two additional statistics, which test whether our model is properly specified. Hansen J's statistic is used to test the overidentifying restrictions of the model. The statistics are insignificant in all four regressions suggesting that the instrumental variables that we used in the regressions are exogenous. Moreover, in the first stage of regressions (not reported here) the instruments explain a significant part of variation in the regulatory variables and the usual F-tests of excluded instruments are significant, which further justify their use as proper instruments.

The second test that we performed consists of including lagged residual into the regression. The aim of the test is to detect serial correlation in the error term, which indicates a misspecification of the model. The lagged residuals are not significant in all four regressions in Table 4, which suggests no serial correlation in the dynamic model.

The next set of results in Table 5 compares the performance of a dynamic model (columns 1 and 2) and a static model (columns 3 and 4). The estimates in the first column of table 5 are the same as in table 4 and table 3. The second column contains results of the same model estimated by OLS. The results in column 1 and column 2 are not statistically different. Accounting for the possible endogeneity of regulation does not alter the results of the dynamic model. This is in strong contrast to the static model. Estimated by OLS the static model shows very different coefficients than the dynamic model. In particular, all regulatory variables seem to have a significant positive impact on infrastructure deployment. This positive effect of regulation disappears in the IV regression in column 4. Inspection of the static model's test statistics also reveals a strong serial correlation in the error term, as evidenced by large and significant coefficients on the lagged residual, and Hansen J's statistics are significantly different from 0. In sum, the static model seems to suffer from omitted variables, which are time persistent and bias the coefficients on regulatory variables. The IV techniques help to alleviate the problem to some extent. In any case, the dynamic model proves superior to the static model.

Table 4: Dynamic Model of Investment: Instrumental Variables (IV) Estimation Results of Alternative Models

Dependent variable: Log(Infr)

	(1)	(2)	(3)	(4)
Log(Infr) (-1)	0.94***	0.95**	0.95***	0.96***
	(0.02)	(0.02)*	(0.01)	(0.01)
Mobile	-0.63	-0.52	-1.09**	-0.80
	(0.49)	(0.56)	(0.48)	(0.51)
Incumb	-0.41***	-0.40***	-0.51***	-0.49***
	(0.08)	(0.08)	(0.13)	(0.14)
Multisec	0.27**	0.31**	0.46***	0.50***
	(0.12)	(0.13)	(0.17)	(0.19)
Log(M&A) * I(M&A)	0.04**	0.04**	0.04*	0.04**
	(0.02)	(0.02)	(0.02)	(0.02)
Log(GDPpc)	0.52**	0.46*	0.22	0.64
	(0.26)	(0.26)	(0.39)	(0.59)
EntryFix[1] * Incumb	-0.02	0.00	-0.16	-0.08
	(0.21)	(0.22)	(0.29)	(0.28)
EntryFix[1] * Entrant	-0.44***	-0.36**	-0.43*	-0.29
	(0.15)	(0.17)	(0.25)	(0.23)
EntryMob[1] * Mobile	0.87	0.73	1.69**	1.23
	(0.82)	(0.91)	(0.78)	(0.81)
Log(Labour)		0.14		0.57
		(0.41)		(1.16)
Debt (-1)		-0.04		-0.02
		(0.05)		(0.06)
Log(PopDens)		-2.50		-4.37
		(1.95)		(4.32)
CompMob			0.36	0.21
			(0.42)	(0.44)
CompFix			0.00	0.00
			(0.00)	(0.00)
Hansen J statistic (Chi-sq(9))	7.31	7.26	6.21	6.23
Residual (-1)	0.05	0.02	-0.02	-0.04
	(0.05)	(0.05)	(0.05)	(0.05)
Observations	730	635	500	445
R-squared	0.96	0.96	0.96	0.96

Notes:
Robust standard errors in parentheses.
The estimates for intercept, country-specific effects and year dummies are not shown.
* Significant at 10%; ** significant at 5%; *** significant at 1% .
[1] Endogenous variables: EntryFix and EntryMobl; Instrumental variables: EntryFixNeighbour, EntryMobNeighbour, Regul, Rile, Europ and interactions thereof.

Table 5: Dynamic vs. Static Model of Investment: Estimation Results

Dependent variable: Log(Infr)

	(1)	(2)	(3)	(4)
	IV	OLS	OLS	IV
Log(Infr) (-1)	0.94***	0.95***		
	(0.02)	(0.01)		
Mobile	-0.63	-0.21	0.77	3.00
	(0.49)	(0.16)	(0.72)	(3.22)
Incumb	-0.41***	-0.42***	0.48	0.82
	(0.08)	(0.09)	(0.56)	(1.24)
Multisec	0.27**	0.25**	2.85***	2.66**
	(0.12)	(0.11)	(0.47)	(1.16)
Log(M&A) * I(M&A)	0.04**	0.03	0.10	0.10*
	(0.02)	(0.02)	(0.07)	(0.06)
Log(GDPpc)	0.52**	0.30	-1.20	-1.38
	(0.26)	(0.21)	(1.01)	(0.92)
EntryFix[1] * Incumb	-0.02	-0.01	3.26***	2.96
	(0.21)	(0.16)	(0.95)	(1.84)
EntryFix[1] * Entrant	-0.44***	-0.50***	1.70***	1.42
	(0.15)	(0.14)	(0.55)	(1.07)
EntryMob[1] * Mobile	0.87	-0.01	3.17***	-0.83
	(0.82)	(0.21)	(1.11)	(5.31)
Hansen J statistic (Chi-sq(9))	7.31	-	-	16.6*
Residual (-1)	0.05	0.06	0.93***	0.94***
	(0.05)	(0.05)	(0.01)	(0.02)
Observations	730	867	1083	935
R-squared	0.96	0.96	0.34	0.32

Notes:
Robust standard errors in parentheses.
The estimates for intercept, country-specific effects and year dummies are not shown.
* Significant at 10%; ** significant at 5%; *** significant at 1% .
[1] Endogenous variables: EntryFix and EntryMob; Instrumental variables: EntryFixNeighbour, EntryMobNeighbour, Regul, Rile, Europ and interactions thereof.

References

Aghion, P., N. Bloom, R. Blundell, R. Griffith and P. Howitt (2005), Competition and Innovation: An Inverted-U Relationship, Quarterly Journal of Economics, 120, pp.701-728

Alesina, A., S. Ardagana, G. Nicoletti and F. Schiantarelli (2005), Regulation and Investment, Journal of the European Economic Association, 3(4), pp.791-825

Baake, P., U. Kamecke and C. Wey (2005), A Regulatory Framework for New and Emerging Markets, Communications & Strategies, 40, pp.123-136

Bourreau, M. and P. Dogan (2005), Unbundling the Local Loop, European Economic Review, 49, pp.173-199

Cave, M. (2004), Making the Ladder of Investment Operational, Unpublished manuscript

Cave, M. and V. Ingo (2003), How Access Pricing and Entry Interact, Telecommunications Policy, 27(10-11), pp.717-727

Chang, H., H. Koski and S.K. Majumdar (2003), Regulation and Investment Behaviour in the Telecommunications Sector: Policies and Patterns in US and Europe, Telecommunications Policy, 27(10-11), pp.677-699

Conway, P. and G. Nicoletti (2006), Product Market Regulation in Non-Manufacturing Sectors in OECD Countries: Measurement and Highlights, OECD Economics Department Working Paper, 530

Crandall, R. W. (2005), Competition and Chaos, Brookings Institution Press, Washington D.C.

Crandall, R. W., A.T. Ingraham and H.J. Singer (2004), Do Unbundling Policies Discourage CLEC Facilities-Based Investment, The B.E. Journals in Economic Analysis & Policy, 4(1)

Crandall, R.W. and H.J. Singer (2003), An Accurate Scorecard of the Telecommunications Act of 1996: Rejoinder to the Phoenix Center Study No. 7, Criterion Economics, L.L.C.

Duso, T. (2005), Lobbying and Regulation in a Political Economy: Evidence from the U.S. Cellular Industry, Public Choice, 122, pp.251–276

Duso, T. and L.-H. Röller (2003), Endogenous Deregulation: Evidence from OECD Countries, Economic Letters, 81, pp.67-71

Eisner, J. and D.E. Lehman (2001), Regulatory Behavior and Competitive Entry, Paper presented at the 14th Annual Western Conference Center for Research in Regulated Industries

Foros, O. (2004), Strategic Investments with Spillovers, Vertical Integration and Foreclosure in the Broadband Access Market, International Journal of Industrial Organization, 22, pp.1-24

Gilbert, R. and D. Newbery (1982), Preemptive Patenting and the Persistence of Monopoly, American Economic Review, 72(3), pp.514-26

Greenstein, S., M. Susan and P.T. Spiller (1995), The Effect of Incentive Regulation on Infrastructure Modernization: Local Exchange Companies' Deployment of Digital Technology, Journal of Economics and Management Strategy, 4(2), pp.187-236

Griffith, R., R. Harrison and H. Simpson (2006): The Link between Product Market Reform, Innovation and EU Macroeconomic Performance, Institute for Fiscal Studies Economic Papers, 243

Gual, J. and F. Trillas (2004), Telecommunications Policies: Determinants and Impact, CEPR Discussion Paper, 4578

Gutiérrez, L.H. (2003), The Effect of Endogenous Regulation on Telecommunications Expansion and Efficiency in Latin America, Journal of Regulatory Economics, 23(3), pp.257-286

Haring, J. And J.H. Rohlfs (2002), The Disincentives for ILEC Broadband Investment Afforded by Unbundling Requirements, Strategic Policy Research Paper

Haring, J., M. Rettle, J.H. Rohlfs and H.M. Shooshan III (2002), UNE Prices and Telecommunications Investment, Strategic Policy Research Paper

Hassett ,K. A., Z. Ivanova, L.J. Kotlikoff (2003), Increased Investment, Lower Prices – the Fruits of Past and Future Telecom Competition, Unpublished manuscript

Hausman, J. and J. Sidak (2005), Did Mandatory Unbundling Achieve its Purpose? Empirical evidence from five countries, Journal of Competition Law and Economics, 1, pp.173-245

Hazlett, T. and C. Bazelon (2005), Regulated Unbundling of Telecommunications Networks: A Stepping Stone to Facilities-Based Competition?, TPRC, Unpublished manuscript

Höffler, F. (2007), Costs and Benefits from Infrastructural Competition: Estimating Welfare Effects from Broadband Access Competition, Unpublished manuscript

Ingraham, S. and J. Sidak (2003), Mandatory Unbundling, UNE-P, and Cost of Equity: Does TELRIC Pricing Increase Risk for Incumbent Local Exchange Carriers?, Criterion Economics, Cambridge, MA.

Jorde, T.M., J.G. Sidak and D.J. Teece (2000), Innovation, Investments, and Unbundling. Yale Journal of Regulation, 17, pp.1-37

Klingemann, H.D., J. Bara, I. Budge, M. Macdonald and A. Volkens (2006), Mapping Policy Preferences II: Estimates for Parties, Electors and Governments in Central and Eastern Europe, European Union and OECD 1990-2003. Oxford: Oxford University Press

Kotakorpi, K. (2006), Access Price Regulation, Investment and Entry in Telecommunications, International Journal of Industrial Organization, 24(5), pp.1013-20

Li, W. and L.C. Xu (2004), The Impact of Privatization and Competition in the Telecommunications Sector around the World, Journal of Law and Economics, 47

London Economics & PricewaterhouseCoopers (2006), An Assessment of the Regulatory Framework for Electronic Communications – Growth and Investment in the EU e-Communications Sector, Report for DG INFSO of EC

Mohnen, P. and L.-H. Röller (2005), Complementarity in Innovation Policy, European Economic Review, 49, pp.1431-1450

Neven, D.J. and L.-H. Röller (2000), The Political Economy of State Aid: Econometric Evidence for the Member States, in: Neven, D.J. and L.-H. Röller (Eds.), The political economy of industrial policy in Europe and the member states. Berlin: Edition Sigma 2000, pp.25-37

Pindyck, R. (2004), Mandatory Unbundling and Irreversible Investment in Telecom Networks, NBER Working Paper, 10287

Renda, A. (2007), Transatlantic Telecom Services: The Pros and the Cons of Convergence, Center for European Policy Studies

Röller, L.-H. and L. Waverman (2001), Telecommunications Infrastructure and Economic Development: A Simultaneous Approach, American Economic Review, 91(4), pp.909-923

Stigler, G. (1971), The Theory of Economic Regulation. The Bell Journal of Economics, 2, pp.3–21

Valletti, T. (2003), The Theory of Access Pricing and Its Linkage with Investment Incentives, Telecommunications Policy, 27(10-11), pp.659-75

Vereda, J. (2007), Unbundling and Incumbent Investment in Quality Upgrades and Cost Reduction, Unpublished manuscript

Vogelsang, I. (2003), Price Regulation of Access to Telecommunications Networks, Journal of Economic Literature, 41(3), pp.830-862

Wallsten, S. (2003), Of Carts and Horses: Regulations and Privatization in Telecommunications Reform, AEI Brookings Joint Center for Regulatory Studies

Wallsten, S. (2005), Broadband Penetration: An Empirical Analysis of State and Federal Policies, AEI Brookings Joint Center for Regulatory Studies, Working Paper 05-12

Wallsten, S. (2006), Broadband and Unbundling Regulations in OECD Countries, AEI Brookings Joint Center for Regulatory Studies, Working Paper 06-16

Waverman, L., M. Meschi, B.Reillier and K.Dasgupta (2007), Access Regulation and Infrastructure Investment in the Telecommunications Sector: an Empirical Investigation, LECG Ltd, Unpublished manuscript

Willig, R. (2003), Investment is Appropriately Stimulated by TELRIC, Unpublished manuscript

Zarakas, W. P., G.A. Woroch, L.V. Wood, D.L. McFadden, N. Ilias and P.C. Liu (2005), Structural Simulation of Facility Sharing: Unbundling Policies and Investment Strategy in Local Exchange Markets, The Brattle Group, Unpublished manuscript

Zenhäusern, P., H. Telser, S. Vaterlaus and P.Mahler (2007), Plaut Economics Regulierungsindex: Regulierungsindex in der Telekommunikation unter besonderer Berücksichtigung der Investitionsanreize, Plaut Economics, Unpublished manuscript

On the Regulation of Next Generation Networks*

Duarte Brito
Universidade Nova de Lisboa†

Pedro Pereira
Autoridade da Concorrência‡

João Vareda
Autoridade da Concorrência‡

February 2008

Abstract

We examine the telecommunications market equilibria when an incumbent firm may invest in a Next Generation Network, and show that the regulator must prove to operators that he is able to commit to his decisions, with the risk of discouraging investment. When the regulator can commit, and in order to induce investment, he must set higher access prices. For intermediate investment costs, the regulator should concede a monopoly to the incumbent, and if investment costs are too high, he should discourage investment. Finally, we show that a two-part tariff allows the regulator to induce investment in the case of no-commitment, but only for low investment costs.

Keywords: Next Generation Networks, Access pricing, Investment

JEL Classification: L43, L51, L96, L98.

*The opinions expressed in this article reflect only the authors' views, and in no way bind the institutions to which they are affiliated. Corresponding author: joao.vareda@autoridadedaconcorrencia.pt

†DCSA, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal.

‡AdC, Rua Laura Alves n^{o}4, 4^o, 1050-038 Lisboa, Portugal.

Introduction

Next Generation Networks. The investment in Next Generation Networks is currently one of the main issues in the telecommunications market debate. There has been a wide discussion about how to regulate the future access to these networks, with incumbent operators claiming the right not to give access to potential entrants, and regulators threatening to force the opening of these infrastructures. Take the example of the dispute between the European Commission, and the German government and Deutsche Telekom, about mandating access to the VDSL network that Deutsche Telekom plans to build in fifty German cities. Deutsche Telekom claimed the right to an access holiday to this future network, and the government offered its support. The European Commission counter-argued that existing ex-ante regulation had to be extended to this network, since the lack of competition in the German market could lead to the re-emergence of a monopoly. A similar discussion is now occurring in Spain.

These investment possibilities, which may involve fibre as close to the home as possible and / or transmission of all data using the IP protocol, poses new problems to regulators, as compared to the previous problem of regulating old monopoly infrastructures, who need to manage the trade-off between the objectives of static and dynamic efficiency. While regulation for static efficiency aims to reduce the market power of incumbent operators, it also reduces the rents on their future investments. Hence, regulators face the difficult task of determining how to encourage operators to invest optimally without lessening competitive intensity too much.

Model and results. In this paper we develop a theoretical model with two operators, an incumbent and an entrant, that compete on prices, by setting two part retail tariffs. The incumbent is a vertically integrated firm and has a network and a retail business, while the entrant only has a retail business, and therefore needs to have access to the incumbent's network to be able to compete in the market. Our model supposes Hotelling competition with consumers buying a subscription to one of the operators, plus minutes of calls, similarly to Biglaiser and DeGraba (2001).

The main contribution of our paper is the determination of the best regulatory practice when the incumbent can invest in a Next Generation Network, namely we discuss about the need to give this incumbent a monopoly on the new network, or if he should continue to give access to the entrant like he was doing with the old network.

We first find that the regulator should try to commit to his policy decisions since in the case he cannot commit and only set the access price to a possible NGN after the investment has been done, there will be no investment, at least with a linear access tariff.

Even with a two-part access tariff there may be no investment by the incumbent when the regulator only sets the access price ex-post, if the investment cost is sufficiently high. This happens because the regulator extracts all the investment rents from the incumbent in order to promote a level playing field competition, given that investment cost is a sunk cost once the incumbent pays it. This result is in the line of the results of Vareda (2007) which shows that when the regulator cannot commit to an unbundling price before investment the incumbent will not have incentives to invest neither in quality upgrades, nor in cost reduction. Indeed, according to Valletti (2003), one of the main issues that must be taken into account on the delineation of regulatory policies is the fact that a regulator should be able to commit to rules over a reasonable time period, i.e., a regulator should try to stabilize his policies in order to show to operators that he can commit to his decisions.

Next we show that when the regulator is able to commit to set access prices before investment, he should not set these at the same level they were before investment, i.e., the access price to the NGN must be higher than the access price to the old network that the entrant was paying previously, in order to allow the incumbent to retain part of his investment rents. Moreover, we find that for a high investment cost it is better to concede a monopoly position to the incumbent, so that every consumer can buy the high quality services at a lower price per minute, and the incumbent obtains the maximum rents from his investment. This is reinforced by the fact that the retail price in this model consists on a two part tariff, where firms set marginal price equal to marginal cost, and extract consumers' surplus through the fixed fee. We further show that it is never optimal to promote a duopoly with the entrant supplying his services through the old network, and the incumbent supplying his services through the new network. Indeed, it is always better to have competition on the NGN, or when this is impossible because the need to obtain investment rents forces the regulator to set a very high access price, it is better to have a monopoly on the NGN since it is, at least, assured that every consumer buys the optimal number of minutes (at marginal cost) despite the higher transportation costs. Given the distortions that the regulator must introduce in order to induce investment by the incumbent, it may even be better to discourage investment in situations where investment is a first-best. This is the same to say that there is under-investment in NGN.

Related literature. The academic literature on regulation has only recently started to address the issues of access pricing and investment. Guthrie (2006) provides a survey on the recent literature about the relationship between infrastructure investment and

the different regulatory regimes, concluding that much has still to be done in this field.

Vareda (2007) studies the incumbent's incentives to invest in quality upgrades and cost reduction when the regulator forces him to unbundle his network, and shows that the regulator should commit to set a lower (higher) unbundling price when cost reduction is relatively less (more) expensive than quality upgrades. Vareda and Hoernig (2007) study the investment of two operators in new infrastructures which allows them to offer new services. Foros (2004) shows that under some conditions the investment by an incumbent in the quality of his network is lower with price regulation since the access price is set equal to marginal cost. Kotakorpi (2006) considers a similar model with vertical differentiation, and obtains similar results.

The remainder of the paper is organized as follows. We describe the model in Section 1. In Section 2 we solve the equilibrium when the regulator cannot commit to the access prices, and in Section 3 we solve the commitment equilibrium. In Section 4 we determine the first best solution, and compare it with private solutions. In Section 5 we comment on the two-part access tariff context. Finally, in Section 6 we conclude. In the Appendix, we present all the proofs.

1 Model

1.1 Environment

We introduce a model of a telecommunications market, where two firms compete on prices. The operators on this market are: the incumbent, which is a vertically integrated firm and has a network and a retail business, and the entrant, which only has a retail business. We assume that the incumbent and the entrant are located on opposite ends of an Hotelling line of length 1. The incumbent is located at 0 and the entrant at 1.

The entrant can only offer his services if he has access to the incumbent's network, for which he must pay an access price. We further assume that the incumbent can build a Next Generation Network (NGN). For instance, he could be supplying his services through a cooper line, and there is the possibility of building a new digital line which allows him to increase the quality of the services offered.

Additionally, we introduce a third party, the regulator, who sets the access prices in order to maximize social welfare. We assume that the access prices are the only instrument available to the regulator, which corresponds closely to the current European practice. Furthermore, we adopt the simplifying assumption of complete information, i.e. the regulator is supposed to have full information about demand and costs.

1.2 Consumers

As we have already referred, we assume a Hotelling type competition in the downstream market: consumers are uniformly distributed along a segment of length 1, facing transportation costs tx to travel the distance x, with $t > 0$. Consumers are otherwise a homogeneous group, meaning that each consumer has the same demand function for the services involved.

Each consumer buys services from only one operator j, with $j \in \{I, E\}$, where I denotes the incumbent and E the entrant, and we assume, as in Biglaiser and DeGraba (2001), that each consumer has a linear demand, given by $y_j = D(p_j) = z - p_j$, where y_j denotes the quantity of minutes purchased to firm j, p_j is the price per minute if the consumer has chosen firm j, and z is a positive parameter.

Denote by $S(p_j) := \frac{(z-p_j)^2}{2}$ a given consumer surplus when purchasing from firm j at unit price p_j. This is gross surplus in the sense that transportation costs and any other fees have still to be deducted. A consumer will only purchase services from an operator if his net surplus is non-negative.

1.3 Firms

We consider that the incumbent produces an input that (i) uses in the production of a final product or (ii) sells to an entrant. Furthermore, we assume that all marginal costs are constant and equal to zero, as well as the fixed costs of operation, and that all other sunk costs have already been incurred by both the entrant and the incumbent.

For each unit produced (correspondent to minutes consumed by his consumers) the entrant pays the wholesale unit price $\alpha \geq 0$, set through regulation, to the incumbent. For now we assume that the access tariff consists only on a variable component. Later, we will consider the possibility of an access tariff composed of a variable and a fixed component.

Firms compete by setting two-part tariffs, denoted by $T_j(y_j) = F_j + p_j y_j$, so that firm j's profits for the whole game are represented by:

$$\pi_I = s_I \left[p_I \left(z - p_I\right) + F_I\right] + s_E \alpha \left(z - p_E\right) \tag{1}$$

$$\pi_E = s_E \left[\left(p_E - \alpha\right)\left(z - p_E\right) + F_E\right] \tag{2}$$

where s_j denotes the market share of operator $j = I, E$.

We will consider that the incumbent can invest in a NGN which allows him to increase the quality of his services. This investment costs $C > 0$, and increases consumer's

demand to $y = (z + \Delta_d) - p_j$. In case of investment, the entrant chooses between asking for access to the old or to the new network depending on the access prices. Given our Hotelling model structure, when both firms compete though the same network there is only horizontal differentiation, but if they compete through different networks, there is also vertical differentiation.

We impose one restriction on the model:

$$z > \frac{4}{3}\sqrt{6t} \tag{3}$$

As we will see below, this assumption on z ensures that the incumbent's profit is increasing in the access price, and implies also that all consumers will have a positive surplus under the different market structures.

1.4 Timing of the game

We will analyze the game starting from the moment that there is a possibility of an investment in a Next Generation Network. However, we assume that, previously to this game, the entrant was asking for access to the old network at a given regulated price. Later we will show that this assumption was indeed the equilibrium of an hypothetical pre-investment game.

We will consider two timings for the investment game: In the case where the sectorial regulator cannot commit to a regulation policy towards the new network, the *no-commitment case*, the game has five stages which unfold as follows. In stage 1, the sectorial regulator sets the access price to the old network. In stage 2, the incumbent makes the investment decision. In stage 3, the sectorial regulator sets the access price to the new network. In stage 4, the entrant decides if he continues to ask for access and to which network, and finally in stage 5 the incumbent and the entrant compete on retail prices. In the case where the sectorial regulator can commit to a regulation policy towards the new network, the *commitment case*, the game has four stages which unfold as follows. In stage 1, the sectorial regulator sets the access prices to the old and the new networks. In stage 2, the incumbent makes the investment decision. In stage 3, the entrant decides if he continues to ask for access and to which network, and in stage 4, the incumbent and the entrant compete on retail prices.

This latter case may seem supported on a strong assumption, but it can be argued that the access prices set by the regulator provide some commitment. For instance, the

regulator can announce that he will set access prices at a certain level for a certain period, for instance until the next review. In this case this timing makes sense: if the incumbent undertakes investments during the same period, he will take as given the access prices set by the regulator. Guthrie (2006) discusses the constraints on the regulator's actions adopted in several countries to prevent him from acting opportunistically. However, in some cases, this may be difficult because of political and/or practical constraints.

1.5 Equilibrium Concept

The sub-game perfect Nash equilibrium is:

(i) a set of retail prices: $(p_I^*, F_I^*, p_E^*, F_E^*) \in \Re_0^+$

(ii) an entry decision by the entrant: $\Phi^* \in$ {ask for access to the NGN; ask for access to the old network; exit the market},

(iii) an investment decision by the incumbent: $\Psi^* \in$ {investment; no investment}

(iv) a set of access prices set by the regulator: $(\alpha_o^*, \alpha_n^*) \in \Re_0^+$

such that:

(E1) $(p_I^*, F_I^*, p_E^*, F_E^*)$ maximize firms profits given the access prices set previously by the regulator, and the market structure;

(E2) Φ^* maximizes entrant's profits given the access prices, and the incumbent's investment decision;

(E3) Ψ^* maximizes the incumbent's profits given the access prices;

(E4) (α_o^*, α_n^*) maximize social welfare.

2 Equilibrium of the No-Commitment Case

In the following sections we will characterize the equilibria of the game for the no-commitment case, which we construct by working backwards. Remember that in this investment timing the regulator only sets the access price to a possible NGN after it has been built.

2.1 Retail price game

First we characterize the equilibrium of the retail price game for three cases: **(i)** the incumbent invests in the NGN and obtains a monopoly position in the retail market, **(ii)** the incumbent invests in the NGN and gives access to it to the entrant, **(iii)** the incumbent invests in the NGN, but only gives access to the old network.

In case **(ii)** the incumbent gives access to the new network at α_n, and in case **(iii)** he gives access to the old network at α_o.

We start with the following Lemma.

Lemma 1 *In equilibrium, firms set the marginal price of the two-part tariff at marginal cost, i.e.* $p_I = 0$ *and* $p_E = \alpha$.

As usual, with two-part tariffs, firms set the variable component of the retail tariff at marginal cost in order to maximize gross consumer surplus, and then try to extract this surplus through the fixed fee component, maximizing their profits.

Given Lemma 1, from now on we will only discuss the determination of fixed fees.

2.1.1 Monopoly

We start to analyze the case where the incumbent obtains a monopoly position in the retail market after investing in a NGN.

Lemma 2 *When in a position of monopoly, and after investing in a NGN, in equilibrium, the incumbent charges the fixed part of the two-part tariff:*

$$F^{Mn} = \frac{(z + \Delta_d)^2}{2} - t. \tag{4}$$

At this fixed charge, the incumbent's profit after investment becomes:

$$\pi^{Mn}(C) = \frac{(z + \Delta_d)^2}{2} - t - C, \tag{5}$$

and total welfare is given by:

$$W^{Mn}(C) = \frac{(z + \Delta_d)^2}{2} - \frac{t}{2} - C. \tag{6}$$

In a context where only the old network is deployed, and the incumbent does not give access to the entrant, the equilibrium is similar to this, but with $\Delta_d = 0$. Therefore, we can obtain incumbent's profit and welfare as special cases of (5), and (6):

$$\pi^{Mo} = \frac{z^2 - 2t}{2}, \qquad W^{Mo} = \frac{z^2 - t}{2}. \tag{7}$$

2.1.2 Duopoly on the New Network

Next, we characterize the case where the NGN is deployed, and the entrant gives access to the new network to the entrant. In this case, both firms face a demand given by $y_j = (z + \Delta_d) - p_j$, and the entrant has a marginal cost given by α_n. We will call this case the "duopoly on the new network", as opposed to the one we will analyze in the next section that we will call "duopoly on the old and new network".

Lemma 3 *When the entrant asks for access to the NGN, in equilibrium, the incumbent and the entrant charge, respectively, the fixed part of the two-part tariff:*

$$F_I^{nn}(\alpha_n) \equiv \begin{cases} (z+\Delta_d)\,\alpha_n + t - \frac{5}{6}\alpha_n^2 & \text{if } \alpha_n < \sqrt{6t} \\ (z+\Delta_d)\,\alpha_n - t - \frac{1}{2}\alpha_n^2 & \text{if } \alpha_n \in \left[\sqrt{6t}, z+\Delta_d\right] \\ \frac{(z+\Delta_d)^2}{2} - t & \text{if } \alpha_n > z+\Delta_d. \end{cases} \tag{8}$$

$$F_E^{nn}(\alpha_n) \equiv \begin{cases} t - \frac{1}{6}\alpha_n^2 & \text{if } \alpha_n < \sqrt{6t} \\ 0 & \text{if } \alpha_n \geq \sqrt{6t}. \end{cases} \tag{9}$$

Given the above fixed charges, in case of investment, the incumbent's and entrant's profits become, respectively:

$$\pi_I^{nn}(\alpha_n, C) = \begin{cases} \frac{1}{2}t + \frac{\alpha_n^4 + 72t(z+\Delta_d)\alpha_n - 60t\alpha_n^2}{72t} - C & \text{if } \alpha_n < \sqrt{6t} \\ (z+\Delta_d)\,\alpha_n - \frac{1}{2}\alpha_n^2 - t - C & \text{if } \alpha_n \in \left[\sqrt{6t}, z+\Delta_d\right] \\ \frac{(z+\Delta_d)^2}{2} - t - C & \text{if } \alpha_n > z+\Delta_d. \end{cases} \tag{10}$$

$$\pi_E^{nn}(\alpha_n) = \begin{cases} \frac{\left(6t-\alpha_n^2\right)^2}{72t} & \text{if } \alpha_n < \sqrt{6t} \\ 0 & \text{if } \alpha_n \geq \sqrt{6t}. \end{cases}, \tag{11}$$

and total welfare is given by:

$$W^{nn}(\alpha_n, C) = \begin{cases} \frac{5\alpha_n^4 + 72t(z+\Delta_d)^2 - 36t\left(t+\alpha_n^2\right)}{144t} - C & \text{if } \alpha_n < \sqrt{6t} \\ \frac{(z+\Delta_d)^2}{2} - \frac{t}{2} - C & \text{if } \alpha_n \geq \sqrt{6t}. \end{cases}. \tag{12}$$

Lemma 4 *In a duopoly on the NGN, the incumbent's (entrant's) profit is increasing (decreasing) in the access price.*

When the access price increases, the incumbent's market share increases since his price per minute (which is given by the marginal cost) becomes relatively more competitive as compared to the entrant's price (given by the access price). In the limit, if the

access price is too high ($\alpha_n \geq \sqrt{6t}$), the entrant is no more able to attract consumers to buy his services. In this case the incumbent is the only operator supplying services in equilibrium, although he only obtains the monopoly profit when the access price is such that the entrant is no more able to contest his position, i.e. when $\alpha_n > z + \Delta_d$. Otherwise, despite the entrant has a zero market share, he is putting some pressure over the incumbent to keep prices lower than monopoly prices.

Note that, as in the monopoly case, the equilibrium where only the old network is deployed, and the incumbent gives access to the entrant, is similar to this, but with $\Delta_d = 0$. Therefore, we can also obtain the incumbent's and the entrant's profit (respectively $\pi_I^{oo}(\alpha_o)$ and $\pi_E^{oo}(\alpha_o)$), and welfare ($W^{oo}(\alpha_o)$) as special cases of (10), (11) and (12):

$$\pi_I^{oo}(\alpha_o) = \begin{cases} \frac{1}{2}t + \frac{\alpha_o^4 + 72tz\alpha_n - 60t\alpha_o^2}{72t} & \text{if } \alpha_o < \sqrt{6t} \\ z\alpha_o - \frac{1}{2}\alpha_o^2 - t & \text{if } \alpha_o \in \left[\sqrt{6t}, z\right] \\ \frac{z^2}{2} - t & \text{if } \alpha_o > z. \end{cases} \tag{13}$$

$$\pi_E^{oo}(\alpha_o) = \begin{cases} \frac{\left(6t - \alpha_o^2\right)^2}{72t} & \text{if } \alpha_o < \sqrt{6t} \\ 0 & \text{if } \alpha_o \geq \sqrt{6t}. \end{cases}, \tag{14}$$

$$W^{oo}(\alpha_o) = \begin{cases} \frac{5\alpha_o^4 + 72tz^2 - 36t\left(t + \alpha_o^2\right)}{144t} & \text{if } \alpha_o < \sqrt{6t} \\ \frac{z^2 - t}{2} & \text{if } \alpha_o \geq \sqrt{6t}. \end{cases}. \tag{15}$$

2.1.3 Duopoly on the Old and New Network

Finally, we analyze the case where the NGN is deployed, and the incumbent gives access to the old network but not to the new one. In this case the demand facing the incumbent is $y_I = (z + \Delta_d) - p_I$, while the demand facing the entrant is $y_E = z - p_E$. The entrant has marginal cost α_o.

Lemma 5 *Define $\alpha_o^{\max} \equiv \sqrt{6t - \Delta_d(2z + \Delta_d)}$. When the entrant asks for access to the old network, in equilibrium, the incumbent and the entrant charge, respectively, the fixed part of the two-part tariff:*

$$F_I^{no}(\alpha_o) \equiv \begin{cases} t - \frac{5\alpha_o^2 - \Delta_d(2z + \Delta_d) - 6\alpha z}{6} & \text{if } \alpha_o < \alpha_o^{\max} \\ \frac{1}{2}(\alpha_o + \Delta_d)(2z - \alpha_o + \Delta_d) - t & \text{if } \alpha_n \in [\alpha_o^{\max}, z] \\ \frac{(z + \Delta_d)^2}{2} - t & \text{if } \alpha_n > z. \end{cases} \tag{16}$$

$$F_E^{no}(\alpha_o) \equiv \begin{cases} t - \frac{\alpha_o^2 + \Delta_d(2z + \Delta_d)}{6} & \text{if } \alpha_n < \alpha_o^{\max} \\ 0 & \text{if } \alpha_n \geq \alpha_o^{\max}. \end{cases} \tag{17}$$

When the entrant asks for access to the old network, and the incumbent supplies his services through a NGN, the incumbent's and entrant's profits, respectively, become:

$$\pi_I^{no}(\alpha_o, C) = \begin{cases} \frac{12t\left(3t+6z\alpha_o-5\alpha_o^2\right)+12\Delta_d(2z+\Delta_d)+\left(\alpha_o^2+\Delta_d^2\right)^2+4z\Delta_d\left(\alpha_o^2+\Delta_d^2+z\Delta_d\right)}{72t} - C & \text{if } \alpha_o < \alpha_o^{\max} \\ \frac{1}{2}(\alpha_o+\Delta_d)(2z-\alpha_o+\Delta_d) - t - C & \text{if } \alpha_n \in [\alpha_o^{\max}, z] \\ \frac{(z+\Delta_d)^2}{2} - t - C & \text{if } \alpha_n > z. \end{cases} \tag{18}$$

$$\pi_E^{no}(\alpha_o) = \begin{cases} \frac{\left(6t-\alpha_o^2-\Delta_d(2z+\Delta_d)\right)^2}{72t} & \text{if } \alpha_o < \alpha_o^{\max} \\ 0 & \text{if } \alpha_n \geq \alpha_o^{\max} \end{cases} \tag{19}$$

and total welfare is given by:

$$W^{no}(\alpha_o, C) = \begin{cases} \frac{5\left(\alpha_o^2+\Delta_d^2\right)^2-36t\left(t-2z\Delta_d+\alpha_o^2-\left(\Delta_d^2+2z^2\right)\right)+20z\Delta_d\left(\Delta_d(z+\Delta_d)+\alpha_o^2\right)}{144t} - C & \text{if } \alpha_n < \alpha_o^{\max} \\ \frac{(z+\Delta_d)^2}{2} - \frac{t}{2} - C & \text{if } \alpha_n \geq \alpha_o^{\max}. \end{cases} \tag{20}$$

Lemma 6 *In a duopoly on the old and new network, the incumbent's (entrant's) profit is increasing (decreasing) in the access price.*

Once again, given that the entrant's profit is decreasing in the access price, since his market share decreases when the access price increases, if the access price is too high ($\alpha_o \geq \alpha_o^{\max}$), the entrant is no more able to attract consumers, and the incumbent becomes a monopolist, although he can only charge monopoly prices for a sufficiently high access price.

Lemma 7 *When $\Delta_d > \overline{\Delta}_d \equiv \sqrt{z^2+6t} - z$, a duopoly on the old and new network is impossible for all $\alpha_o \geq 0$.*

When the new technology is so much better than the old one, the entrant can never obtain a positive market share by supplying his services through the old technology, even for a zero access price. In this case, the regulator can only promote a duopoly through the new network.

2.2 Entry decision

Given the equilibrium profits the entrant can obtain in each of the possible retail stage sub-games, we will now solve stage 4 of the game, i.e. we will determine for each pair of access prices (α_n, α_o) if the entrant prefers to continue to ask for access to the incumbent's old network, or if he prefers to start asking for access to the NGN, or if he exits the market.

The next Lemma presents the optimal decision by the entrant. We assume that in case of indifference between asking for access or exit the market, the entrant exits the market, and in case of indifference between continuing to ask for access to the old network and asking for access to the new network, he will prefer the former.

Lemma 8 *Define* $\alpha_n^{\min}(\alpha_o) \equiv \min\left\{\sqrt{6t}, \sqrt{\alpha_o^2 + \Delta_d(2z + \Delta_d)}\right\}$. *If the incumbent has built a new network and* $\Delta_d < \overline{\Delta}_d$, *the entrant:*

i) asks for access to the new network if $\alpha_n < \alpha_n^{\min}(\alpha_o)$,

ii) continues to ask for access to the old network if $\alpha_o < \alpha_o^{\max}$ *and* $\alpha_n \geq \alpha_n^{\min}(\alpha_o)$,

iii) exits the market if $\alpha_o \geq \alpha_o^{\max}$ *and* $\alpha_n \geq \alpha_n^{\min}(\alpha_o)$.

If $\Delta_d \geq \overline{\Delta}_d$, *then the entrant asks for access to the new network if* $\alpha_n < \sqrt{6t}$, *and exits the market otherwise.*

When the access price to the new network is low as compared to the access price to the old network, than the entrant prefers to ask for access to the latter. Note that this is true even for some $\alpha_n > \alpha_o$. In fact, the entrant may prefer to pay a higher access price to the new network as compared to the old one, because he can offer a higher quality service through the new network. On the contrary, when the difference between the access prices is too high in favour of the old network, then the entrant prefers to continue to ask for access to this network, despite the lower quality of the services he will be able to offer. If both access prices are too high, then the entrant exits the market.

When the NGN represents a big improvement to the quality of the services supplied, then the entrant will never consider to ask for access to the old network since he would be unable to attract any consumer.

Lemma 9 *If the incumbent does not invest in a new network, the entrant continues to asks for access to the old network if* $\alpha_o < \sqrt{6t}$, *and exits the market otherwise.*

If the incumbent does not invest, then the entrant's decision is simply to stay in the market if he is able to obtain a positive market share in a duopoly on the old network, and exit the market otherwise.

2.3 Regulation of the new network

In this section we characterize the regulator's optimal decision about the access price to the new network, assuming that it has already been built by the incumbent. The regulator acts as a second-mover, in the sense that he decides the access price only after observing the incumbent's investment decision, and thus he considers investment costs as sunk costs. The regulator then chooses α_n which maximizes $W^{nn}(\alpha_n, C)$, represented as in Figure 1, without any investment restriction.

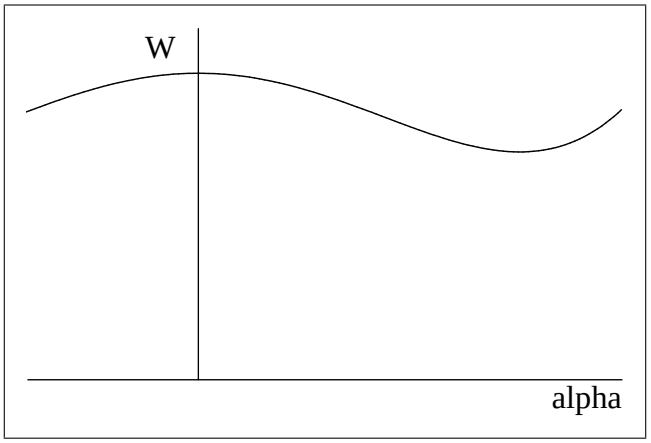


Figure 1

According to the previous Figure, the candidates to welfare maximizers are an access price equal to zero, or an access price at which the entrant does not want to ask for access any more.

Lemma 10 *When the incumbent has invested in a NGN, the regulator sets ex-post* $\alpha_n = 0$.

Lemma 10's result is intuitive. Indeed, from Lemma 1, we know that $\alpha_n = 0$ induces the entrant to charge the lowest retail price per minute, and at the same time minimizes the transportation cost since at this access price firms share the market equally. Moreover, at this access price the entrant prefers to ask for access to the NGN, and therefore all consumers buy the services through the new network.

2.4 Investment decision

Following the timing of the game backwards, we will now characterize the optimal investment decision by the incumbent.

Lemma 11 *When the regulator sets the access price to a NGN ex-post, in equilibrium, the incumbent does not invest in the new network.*

The incumbent does not invest in the NGN since he foresees that the regulator will ex-post extract all the rents from his investment by setting $\alpha_n = 0$. This is similar to the result of Vareda (2007) where the incumbent also does not invest neither in quality upgrades nor in cost reduction when the regulator is not able to commit to an unbundling price. This happens because in an Hotelling model, when firms face the same marginal cost, their profit depends only on the differences in the services' quality, and not on the absolute value of qualities. By this way, if both firms benefit from the same increment in quality, the profit level remains constant, and consumers take all the benefits from investment.

If we had considered different spillover effects of investment, i.e. if the dimension of the investment effect depended on the ability of each operator to transform input to output such that we would have the incumbent offering higher quality services, this would not be truth. In fact, if the incumbent could offer higher quality services after investing, he could appropriate some of his investment gains, and would then have some incentives to invest.[1] This would also be true if the market was not totally covered, so that the increase in the services quality would bring new consumers to the market, increasing firms' profit. In a later section, we will analyze the case of an access tariff composed of a variable and a fixed component. Also in this case, it is possible to observe investment, in equilibrium, if the investment cost is sufficiently low.

2.5 Regulation of the old network

Finally, we solve the regulator's problem at the first stage, i.e. when he takes the decision about the access price to the old network. Given that the regulator foresees that there will be no investment in a NGN, he sets the access price to the old network taking this into consideration. Therefore, as the socially optimal access price when firms compete through the same network does not depend on the quality of the network, it is the same as under competition on a NGN.

Lemma 12 *When the regulator sets the access price to the NGN ex-post, he should set the access price to the old network at* $\alpha_o = 0$.

[1] This is the assumption on Foros (2004) and Kotakorpi (2006).

Given this Lemma, in the hypothetical pre-investment game, the socially optimal access price would also be $\alpha_o = 0$, and the entrant would be asking for access to the old network, and firms sharing the market equally.

2.6 Equilibrium

Having solve all the five stages of the no-commitment game, we are now able to summarize its equilibrium:

Proposition 13 *The equilibrium of the no-commitment game is* $\alpha_o^* = 0$; $\Psi^* = \{$*no investment*$\}$; $\alpha_n^* = 0$; $\Phi^* = \{$*asks for access to the old network*$\}$; $F_I^* = F_E^* = t$, *and* $p_I^* = p_E^* = 0$.

In a context where the regulator cannot commit and the entrant is able to offer the same quality of services as the incumbent by having access to a possible NGN that the latter could build, the possibility of investment does not change the equilibrium of as compared to the pre-investment game. It is as if there was no possibility of investment, and the market does not develop to a Next Generation Network.

This can be seen as an alert to regulators, which by looking only to the static efficiency of the market, risk to discourage many fundamental investments. Next, we will analyze what happens in the case where the regulator adopts a more stable practice, and convinces operators that he is able to commit to the access prices set before investment.

3 Equilibrium of the Commitment Case

In this section, we will determine the equilibrium of the game for the commitment case, which we construct by working backwards. Since the retail price game is identical to that of the no-commitment case, as well as the entrant's decision about which network to ask for access, they are omitted and we proceed directly to stage 3: the investment decision.

3.1 Investment Decision

First, we will characterize the optimal investment decision by the incumbent, given the access prices to the new and old network set previously by the regulator. We must do this for all possible set of access prices (α_o, α_n) since the regulator may find it optimal

to promote a duopoly on the new network, or a duopoly on the old and new network, or even a monopoly on the new network, at the regulation stage.

Duopoly on the new network

This scenario can only take place when $\alpha_n < \alpha_n^{\max}(\alpha_o)$. We have to consider two sub-cases: the first where in case of no investment the entrant continues to ask for access to the old network, and the second where in case of no investment the entrant exits the market since the access price to the old network becomes too high. Given these, the incumbent invests if and only if:

$$\pi_I^{nn}(\alpha_n, C) \geq \begin{cases} \pi_I^{oo}(\alpha_o) & if \quad \alpha_o < \sqrt{6t} \\ \pi^{Mo} & if \quad \alpha_o \geq \sqrt{6t} \end{cases} \tag{21}$$

Since the incumbent's profit is increasing in the access price, for $\alpha_o < \sqrt{6t}$ this condition is equivalent to $\alpha_n \geq \alpha_n^{inv}(\alpha_o, C)$, where $\alpha_n^{inv}(\alpha_o, C)$ is defined by $\pi_I^{nn}\left(\alpha_n^{inv}, C\right) = \pi_I^{oo}(\alpha_o)$. Note that for $\alpha_o < \sqrt{6t}$, $\frac{d\alpha_n^{inv}(\alpha_o,C)}{dC} > 0$ and $\frac{d\alpha_n^{inv}(\alpha_o,C)}{d\alpha_o} > 0$, since $\frac{\partial \pi_I^{nn}(\alpha_n,C)}{\partial C} < 0$ and $\frac{\partial \pi_I^{oo}(\alpha_o)}{\partial \alpha_o} > 0$, which means that the higher is α_o and C, the higher must be α_n for the incumbent to invest in a NGN. When $\alpha_o \geq \sqrt{6t}$, then $\alpha_n^{inv}(\alpha_o, C)$ is defined by $\pi_I^{nn}\left(\alpha_n^{inv}, C\right) = \pi^{Mo}$. In this case, $\alpha_n^{inv}(\alpha_o, C)$ is continuous in α_o and increasing in C.

According to our analysis in Lemma 11, we necessarily have $\alpha_n^{inv}(\alpha_o, C) > 0$, i.e. the incumbent will not invest if the access price to the new network is zero. We further know that there is a $\widetilde{C}(\alpha_o)$, which is the highest C such that (21) is possible for some $\alpha_n < \alpha_n^{\max}(\alpha_o)$, i.e. above this $\widetilde{C}(\alpha_o)$ we have $\alpha_n^{inv}(\alpha_o, C) \geq \alpha_n^{\max}(\alpha_o)$, or $\alpha_n^{inv}(\alpha_o, C)$ can no more be defined. In this case, the incumbent will not invest for any α_n.

Duopoly on the old and new network

This scenario can only take place if $\alpha_o < \alpha_o^{\max}$ and $\alpha_n \geq \alpha_n^{\max}(\alpha_o)$, i.e. if the access prices are such that the entrant prefers to continue to ask for access to the old network, even when the incumbent invests in a NGN. In this case the incumbent invests if and only if:

$$\pi_I^{no}(\alpha_o, C) \geq \pi_I^{oo}(\alpha_o). \tag{22}$$

This condition can be translated in $\alpha_o \geq \alpha_o^{inv}(C)$, whenever $C > \frac{\Delta_d(2z+\Delta_d)\left(12t+2z\Delta_d+\Delta_d^2\right)}{72t}$, where $\alpha_o^{inv}(C)$ is defined by $\pi_I^{no}\left(\alpha_o^{inv}, C\right) = \pi_I^{oo}\left(\alpha_o^{inv}\right)$. For a lower C, the incumbent invests for all $\alpha_o \geq 0$. For a very high C, once again, investment is not possible since $\pi_I^{no}(\alpha_o, C) < \pi_I^{oo}(\alpha_o)$ for all $\alpha_o < \alpha_o^{\max}$.

Note that the scenario where in case of no investment the entrant exits the market is impossible since the condition for the entrant to exit the market is stronger in a duopoly

on the old and new network, than in a duopoly on the old network. Thus, if the entrant prefers to exit the market in case of no investment, it will also exit the market in case of investment.

If $\Delta_d > \overline{\Delta}_d$ the entrant will never ask for access to the old network when the entrant builds a NGN because the difference in qualities is too high.

Monopoly on the new network

As in the duopoly on the new network scenario, we also have to consider two sub-cases: the first where in case of no investment the entrant continues to ask for access to the old network, and the second where whatever the investment decision by the incumbent the entrant exits the market.

For $\alpha_o \in \left[\alpha_o^{\max}, \sqrt{6t}\right]$ and $\alpha_n \geq \alpha_n^{\max}(\alpha_o)$, the incumbent invests if and only if:

$$\pi^{Mn}(C) \geq \pi_I^{oo}(\alpha_o). \tag{23}$$

In this case the incumbent obtains a monopoly position in case of investment, but not in case of no investment. For a sufficiently low C, we have $\pi^{Mn}(C) \geq \pi_I^{oo}(\alpha_o)$ for all α_o, while for a sufficiently high C the incumbent will not invest for any $\alpha_o \geq 0$.

For $\alpha_o \geq \sqrt{6t}$ and $\alpha_n \geq \alpha_n^{\max}(\alpha_o)$ the incumbent invests if and only if:

$$\pi^{Mn}(C) \geq \pi^{Mo}. \tag{24}$$

In this case, the incumbent invests when the increase in monopoly profit due to investment is higher than investment cost, i.e. when $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$.

3.2 Regulation of the New and Old Network

We will now characterize the regulator's optimal decision about the access prices to the old and new networks. We will often refer to the fact that the investment is or is not optimal from a social welfare point of view, but this will always be in the perspective that the regulator can only control the access prices to maximize social welfare, being the investment decision left to the incumbent.

According to what we have seen in the previous section, the investment game may end up in a sub-game where there is no investment in a NGN, or in a sub-game where the incumbent invests in a NGN, depending on the access prices (α_o, α_n) and C. The regulator is able to choose (α_o, α_n), and therefore he has some ability to conduct the game to an investment sub-game or to a no investment sub-game. However, when investment

cost is too high, the game continues to a no investment sub-game whatever the regulator does, and in this case his optimal decision is trivial, since he just need to regulate access to the old network as he was doing before the investment possibility.

When investment is cost is sufficiently low, so that the regulator has the faculty of inducing to both sub-games, we need to determine the access prices which maximize welfare on each sub-game, subject to the condition that these must be such that, in equilibrium, the game continues to that sub-game, and then compare welfare levels. Note that in this case it can be socially optimal to set access prices which discourage investment, when investment cost is too high as compared to the increase in the quality of services, or when the distortions introduced on the access price to induce investment by the incumbent are so high, that competition on the old network is socially preferred.

First, we consider the trivial case of determining the access prices which maximize welfare on the no-investment sub-game. If there is no investment, both firms will compete through the old network, and we have already shown that when firms compete through the same network, welfare is maximized at $\alpha_o = 0$. Although α_n does not change welfare, the regulator must set it at a level such that the incumbent does not want to invest, i.e. $\alpha_n < \alpha_n^{inv}(0, C)$.

Lemma 14 *The access prices which maximize welfare at the no investment sub-game are* $\alpha_o = 0$, *and any* $\alpha_n < \alpha_n^{inv}(0, C)$.

Next we determine the access prices which maximize welfare at the investment sub-game. This is the most interesting case, as it will imply that the regulator has to introduce some distortions on the access prices to the old and new network. In fact, according to the previous section, in order not to discourage investment, the regulator must set an $\alpha_n > 0$. Because of this distortion, it is not clear if the access prices which maximize welfare at this sub-game are such that the entrant asks for access to the new or to the old network. It may even be optimal to have a monopoly on a NGN.

Lemma 15 *It is never optimal to set access prices such that at investment sub-game the entrant prefers to continue to have access to the old network.*

A duopoly on the old and new network implies that not all consumers buy the services from a NGN, since the entrant continues to offer his services through the old network. According to the previous Lemma this will never be optimal despite the lower

transportation costs as compared to the monopoly context, and the lower distortions on the access price that are needed to induce investment as compared to a duopoly on the new network. We can then ignore this possibility in the determination of the socially optimal access prices.

Given that we do are not able to obtain an expression for $\alpha_n^{inv}(\alpha_o, C)$, we need to assume that it is a well defined function.

Lemma 16 *Define* $\widetilde{\alpha}_o(C)$, *as the lowest* α_o *such that* $\alpha_n^{inv}(\alpha_o, C)$ *exists and is lower than* $\alpha_n^{\max}(C)$ *for a given* C, $\widehat{C}_1$ *by* $W^{Mn}\left(\widehat{C}_1\right) = W^{nn}\left(\alpha_n^{inv}\left(\widetilde{\alpha}_o\left(\widehat{C}_1\right), \widehat{C}_1\right), \widehat{C}_1\right)$, *and* $\widehat{C}_2 \equiv \max\left\{\widetilde{C}(\alpha_o)\right\}$.

The access prices which maximize welfare at the investment sub-game are:

i) $\alpha_o = \widetilde{\alpha}_o(C)$ *and* $\alpha_n = \alpha_n^{inv}(\widetilde{\alpha}_o(C), C)$ *if* $C \leq \min\left\{\widehat{C}_1, \widehat{C}_2\right\}$ *(duopoly on the new network);*

ii) $\alpha_o \geq \alpha_o^{\max}$ *and* $\alpha_n \geq \sqrt{6t}$ *if* $C > \min\left\{\widehat{C}_1, \widehat{C}_2\right\}$ *(monopoly on the new network).*

According to this Lemma, for low values of C, welfare at the investment sub-game is maximized with a duopoly on the new network since the distortions introduced on the access price are relatively low. For high values of C, it is welfare maximizing to have a monopoly on the NGN. In this case, we have higher transportation costs, but the retail price per minute paid by that every consumer is the efficient one.

When $\Delta_d \geq \overline{\Delta}_d$, the regulator can set $\alpha_o = 0$, since the entrant will never consider to ask for access to the old network when the incumbent is supplying his services through a NGN. However, for $\Delta_d < \overline{\Delta}_d$, i.e. when the increase in the services' quality is not too high, the regulator must set $\alpha_o > 0$, despite there will no operations through the old network, in order to discourage the entrant from asking for access to the old network.

We can then conclude that the higher is C, the higher are the distortions the regulator needs to introduce in order to induce to an investment sub-game. This implies that there will be a $\overline{C}$,above which it is socially optimal to set access prices that conduct to a no-investment sub-game.

Proposition 17 *There is a* $\overline{C} \in \left(\frac{\Delta_d(2z+\Delta_d)}{2} - \frac{1}{4}t, \frac{\Delta_d(2z+\Delta_d)}{2}\right)$ *such that*

- for $C \leq \overline{C}$, *the regulator should set access prices according to Lemma 16,*

- for $C > \overline{C}$, *the regulator should set access prices according to Lemma 14.*

The distortions introduced in the access prices are higher the lower is the increase in the quality of the new network, since it is more difficult to give incentives for an

investment. As as consequence, $\overline{C}$ is higher the lower is Δ_d, i.e., the investment cost above which the regulator starts discouraging investment is higher the lower are the gains this investment brings in terms of the services' quality.

3.3 Equilibrium

Having solve all the four stages of the commitment case, we are now able to summarize its equilibrium:

Proposition 18 *The equilibrium of the no-commitment game is given by:*
i) if $C \leq min\left\{\overline{C}, \widehat{C}_1, \widehat{C}_2\right\}$ *it is* $\alpha_o^* = \alpha_o^{\min}(C)$ *and* $\alpha_n^* = \widehat{\alpha}_n\left(\alpha_o^{\min}(C), C\right)$; $\Psi^* = \{investment\}$; $\Phi^* = \{ask\ for\ access\ to\ the\ NGN\}$; $F_I^* = F_I^{nn}(\alpha_n^*)$, $F_E^* = F_E^{nn}(\alpha_n^*)$, *and* $p_I^* = 0, p_E^* = \alpha_n^*$.
ii) if $C \in \left(min\left\{\overline{C}, \widehat{C}_1, \widehat{C}_2\right\}, \overline{C}\right]$ *it is* $\alpha_o \geq \alpha_o^{\max}$ *and* $\alpha_n \geq \sqrt{6t}$ $\Psi^* = \{investment\}$; $\Phi^* = \{exit\ the\ market\}$; $F_I^* = F_I^{Mn}$, *and* $p_I^* = 0$
iii) if $C > \overline{C}$ *it is* $\alpha_o^* = 0$ *and* $\alpha_n^* < \alpha_n^{inv}(0, C)$; $\Psi^* = \{no\ investment\}$; $\Phi^* = \{ask\ for\ access\ to\ the\ old\ network\}$; $F_I^* = F_E^* = t$, *and* $p_I^* = p_E^* = 0$.

When the investment cost is low the regulator should set access prices such that the incumbent invests in a NGN, and the entrant asks for access to this new network. The access price to the NGN is higher than the access price the entrant was paying previously to the appearance of the NGN, in order to give incentives for the investment, and therefore, the incumbent's market share on the NGN is higher than one half of the market.

For intermediate values of C, and in order assure that every consumer buys the optimal amount of minutes, it may be welfare maximizing to induce to the exit of the entrant, and promote a monopoly on a NGN.

Finally, for a sufficiently high C the regulator should discourage investment because the gains in terms of quality it might bring are lower as compared to its cost. In this case the market stays as it was before the appearance of the investment opportunity.

4 First-best analysis

4.1 First best investment

In this section we will determine when it is a first best to have an investment in a NGN, assuming that the regulator controls not only the access prices but also the investment decision and retail prices.

The retail prices which maximize social welfare are, as usual, determined by the condition p =marginal cost, which in this model case is equal to zero. To maximize welfare we also need to minimize the transportation costs, which is achieved when the indifferent consumer is located at 1/2 of the Hotelling line. This happens when the fixed fee payments are equal for both firms ($F = F_I = F_E$), and firms offer the same quality of services. The fixed payments are in this case only a transfer between firms and consumers, and do not influence welfare, at least until the indifferent consumer obtains a non-negative surplus, i.e., until $F \leq \frac{z^2}{2} - \frac{t}{2}$.

Given this, in case of no investment in a new network, the maximum welfare level is:

$$W_O = \frac{z^2}{2} - \frac{1}{4}t, \tag{25}$$

and in case of investment in a NGN, the maximum welfare is equal to:

$$W_N = \frac{(z + \Delta_d)^2}{2} - \frac{1}{4}t - C. \tag{26}$$

Proposition 19 *Investment is a first best socially optimal when:*

$$\frac{\Delta_d (2z + \Delta_d)}{2} \geq C. \tag{27}$$

4.2 Private and socially optimal investment

In section 3.3 we have shown that for $C > \frac{\Delta_d(2z+\Delta_d)}{2}$, the regulator sets $\alpha_o = 0$ and $\alpha_n = 0$, which discourages investment. In the previous section we have also shown that for the same levels of C, investment is not a first best. Therefore, we conclude that there will never be over-investment in a NGN since the regulator is able to discourage investment and maximize welfare whenever investment is not a first best.

However, for $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$, and when the regulator is able to commit to his decisions, he may prefer to discourage investment on a NGN, despite it is a first best to have the investment done, because of the distortions he needs to introduce to induce investment. Hence, and according to what we have seen previously, for $C \in \left(\overline{C}, \frac{\Delta_d(2z+\Delta_d)}{2}\right]$, it is a first-best to have investment on a new network, but the regulator sets access prices which discourage investment. In this case, there is under-investment on NGN.

Corollary 20 *When the regulator is able to commit to access prices set ex-ante, there is never over investment on a NGN, but there is under investment for* $C \in \left(\overline{C}, \frac{\Delta_d(2z+\Delta_d)}{2}\right]$. *For* $C \leq \overline{C}$, *investment is, at the same time, socially optimal and the first best.*

When the regulator is not able to commit to access prices set ex-ante there will never be investment, and therefore there is under investment as compared to first best for every $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$.

Corollary 21 *When the regulator only sets the access price to the NGN ex-post, there is under investment for* $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$, *and there is never over investment.*

5 Two-Part Tariffs

In this section, we will consider the case where the access price paid by the entrant to the incumbent is composed of a variable component, which depends on the minutes the entrant's consumers buy, and a fixed component $P \geq 0$, which is independent of the number of minutes and of the number of consumers the entrant obtains. This fixed component does not change the nature of retail stage equilibrium since it is a simple transference from the entrant to the incumbent. It will only influence the decision of exiting the market in the sense that from the profits calculated above we have to discount/sum the fixed payment.

Again, we assume that this access tariff is set by the regulator, who maximizes social welfare. The regulator now gains an additional instrument to control, and therefore he will be able to achieve higher levels of welfare as compared to the previous sections. The regulator may even be able to achieve the first best welfare level for more values of the parameters.

Defining (α_o, P_o) as the access tariff to the old network, and (α_n, P_n) the access tariff to the new network, we find:

Proposition 22 *If* $\Delta_d \geq \overline{\Delta}_d$ *and* $t > 2C$ *the regulator achieves the first best by setting:*

- $\alpha_o = 0$, $\alpha_n = 0$, $P_o = 0$ *and any* $P_n \in \left[C, \frac{1}{2}t\right)$ *if* $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$
- $\alpha_o = 0$, $\alpha_n = 0$, $P_o = 0$ *and* $P_n = 0$ *if* $C > \frac{\Delta_d(2z+\Delta_d)}{2}$.

Proposition 23 *Define* $\underline{\alpha}_0$ *as the* $\arg\min_{\alpha_o \in [0,\alpha_o^{\max}]} \left\{\pi_I^{oo}(\alpha_o) + \pi_E^{no}(\alpha_o)\right\}$

If $\Delta_d < \overline{\Delta}_d$, *and* $t > C + \pi_I^{oo}(\underline{\alpha}_0) + \pi_E^{no}(\underline{\alpha}_0)$ *the regulator achieves the first best by setting:*

- $\alpha_o = \underline{\alpha}_0$; $\alpha_n = 0$ *and any* (P_n, P_o) *such that* $P_n - P_o \geq \pi_I^{oo}(\underline{\alpha}_0) - \frac{1}{2}t + C$, $P_o < \pi_E^{oo}(\underline{\alpha}_0)$, *and* $P_n < \frac{1}{2}t - \max\left\{\pi_E^{no}(\underline{\alpha}_0) - P_o, 0\right\}$ *if* $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$
- $\alpha_o = 0$, $\alpha_n = 0$, $P_o = 0$ *and* $P_n = 0$ *if* $C > \frac{\Delta_d(2z+\Delta_d)}{2}$.

Under the above conditions there will be no cost of commitment by the regulator since the socially optimal access tariff to the NGN set before investment is also socially optimal after investment. Indeed, whenever we are under the above conditions and $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$ the regulator is able to induce investment without distorting the variable component of the access price, and therefore the number of minutes of calls chosen by each subscriber will be the socially optimal one. The equilibrium will then be a duopoly on the new network with firms charging $p_j^* = 0$ and $F_j^* = t$, with $j = I, E$.

If $C > \frac{\Delta_d(2z+\Delta_d)}{2}$, as with the linear tariff, the regulator sets access prices such that it is not optimal for the incumbent to invest, and the equilibrium will be a duopoly on the old network, with firms charging again $p_j^* = 0$ and $F_j^* = t$.

The regulator will not have any incentives to change his decision after observing the incumbent's investment. This is true because in order to give incentives to invest the regulator can now compensate the incumbent through the fixed component of the access tariff, and not by distorting the variable component, and thus the optimal price after investment is equal to the one set before. Of course, when $\Delta_d \geq \overline{\Delta}_d$ it becomes easier to achieve the first best since the option of asking for access to the old network when the incumbent has invested is not profitable to the entrant, and thus the regulator can set $P_o = 0$ and $\alpha_o = 0$, so that the gains from investing are higher. When $\Delta_d < \overline{\Delta}_d$ the regulator may be forced to set a $P_o > 0$ and $\alpha_o > 0$, in order to discourage a duopoly on the old and new network, which makes it more difficult to give incentives for investment since that increases the profits in case of no investment.

If $t < 2C$ for $\Delta_d \geq \overline{\Delta}_d$ or $t < C + \pi_I^{oo}(\underline{\alpha}_0) + \pi_E^{no}(\underline{\alpha}_0)$ for $\Delta_d < \overline{\Delta}_d$, i.e., if the investment cost is too high, the regulator is unable to set fixed fees that allow the incumbent to obtain enough rents from his investment, and therefore he must introduce distortions on the variable component of the access tariff in order to induce to an investment sub-game. The results will then come similar to the previous sections.

Indeed, with no-commitment there will be no investment by the incumbent, since he foresees that the regulator will change the variable component of the access tariff to zero after observing the investment, and thus the incumbent will not obtain the necessary rents to invest. The equilibrium will then be equal to the linear tariff with no-commitment case (see Proposition 13). In case of commitment, the regulator may prefer to promote a duopoly on the new network with $\alpha_n > 0$, or a monopoly on the new network, or even to discourage investment, for $C \leq \frac{\Delta_d(2z+\Delta_d)}{2}$. The distortions introduced are similar to the ones obtained with the linear tariff, although of a lower level since the regulator will also use the fixed component to induce to the investment

sub-game, and therefore the equilibrium will also be similar.

6 Conclusion

This paper tries to determine what are the best regulatory practices in the telecommunications market in the context of investments in Next Generation Networks. Contrary to previous regulatory practices, where there was already a network operating in the market, now regulators should take into consideration that they must regulate access to networks that may only be built in the future. Therefore, they have to consider, not only static objectives, but also dynamic objectives.

We show that the regulator should make all the efforts to prove to market operators that he is able to commit to his decisions, with the risk of discouraging investment by the incumbent in NGN. Even when he is able to commit, and in order to induce investment, optimal regulation involves introducing some distortions on the access prices. For an intermediate investment cost, the regulator should concede a monopoly position to the incumbent, so that he can retain his investment gains. If the investment cost is sufficiently high, optimal regulation involves discouraging investment, even in situations where investment would be a first-best.

Finally, we show that the possibility of setting a two-part access tariff does allow the regulator to induce investment in case of no-commitment, but only for low investment costs. For high investment costs, the results come similar to the linear access tariff.

Appendix

Lemma 1: See Biglaiser and DeGraba (2001). ■

Lemma 2: We first analyze the case where the entrant is a monopolist in the retail market. Consumers purchase if and only if

$$\frac{(z+\Delta_d)^2}{2} - tx - F_I > 0 \Leftrightarrow x < -\frac{1}{t}\left(F_I - \frac{1}{2}(z+\Delta_d)^2\right).$$

Assuming an interior solution, the profit maximizing price and respective profits are

$$\begin{aligned} F_I &= \frac{(z+\Delta_d)^2}{4} \\ \pi_I &= \frac{(z+\Delta_d)^4}{16t}. \end{aligned}$$

However, we do not have an interior solution since, given our assumption (3):

$$x = \frac{(z+\Delta_d)^2}{4t} > 1.$$

In this case, the optimal fixed charge and profits are:

$$F_I = \pi_I = \frac{(z+\Delta_d)^2}{2} - t.$$

Consumer surplus and welfare are equal to

$$\begin{aligned} CS &= \frac{\frac{(z+\Delta_d)^2}{2} - \frac{(z+\Delta_d)^2}{2} + t + 0}{2} = \frac{1}{2}t \\ W &= \frac{(z+\Delta_d)^2}{2} - \frac{1}{2}t. \end{aligned}$$

■

Lemma 3 and 5: With respect to duopoly equilibrium, and to avoid the multiplicity of cases, we assume that firm $j = I, E$ faces demand $y_j = (z + \Delta_j) - p_j$ with $\Delta_j \in \{0, \Delta_d\}$. Additionally the entrant has costs $\alpha \in \{\alpha_o, \alpha_n\}$. Let $D := (\Delta_I - \Delta_E)(2z + \Delta_I + \Delta_E)$. Clearly, $D \in \{0, \Delta_d(2z + \Delta_d)\}$, with $D = 0$ when both firm use the same (new or old) network.

We start by finding the consumer who is indifferent between buying from the incumbent or from the entrant:

$$\begin{aligned} \frac{(z+\Delta_I)^2}{2} - tx - F_I &= \frac{(z+\Delta_E-\alpha)^2}{2} - t(1-x) - F_E \Leftrightarrow \\ x(F_I, F_E, \Delta_I, \Delta_E, \alpha; z) &= \left(\frac{1}{2} - \frac{F_I - F_E}{2t} - \frac{(z-\alpha+\Delta_E)^2 - (z+\Delta_I)^2}{4t}\right). \end{aligned}$$

with $\alpha < z$.

Given this indifferent consumer, and the fact that $p_I = 0$ and $p_E = \alpha$, profit functions, excluding investment costs, become:

$$\begin{aligned}\pi_I &= F_I x\left(F_I, F_E, \Delta_I, \Delta_E, \alpha; z\right) + \alpha\left(z + \Delta_E - \alpha\right)\left(1 - x\left(F_I, F_E, \Delta_I, \Delta_E, \alpha; z\right)\right)\\ \pi_E &= F_E(1 - x\left(F_I, F_E, \Delta_I, \Delta_E, \alpha; z\right)).\end{aligned}$$

Maximizing each profit function with respect to the fixed fee, we find:

$$\begin{aligned}F_I^* &= \left(t + z\alpha - \frac{5}{6}\alpha^2 + \alpha\Delta_E + \frac{1}{6}D\right)\\ F_E^* &= \left(t - \frac{1}{6}\alpha^2 - \frac{1}{6}D\right)\end{aligned}$$

The indifferent consumer is given by

$$x^* = \left(\frac{1}{2} + \frac{1}{12t}\left(D + \alpha^2\right)\right),$$

with $\alpha \leq \sqrt{6t - D}$, for each D.

Equilibrium profits are then:

$$\begin{aligned}\pi_I^* &= \frac{\left(36t^2 + \alpha^4 - 60t\alpha^2\right) + 72\alpha t\left(z + \Delta_E\right) + D\left(12t + D + 2\alpha^2\right)}{72t}\\ \pi_E^* &= \frac{\left(6t - \alpha^2 - D\right)^2}{72t}.\end{aligned}$$

Consumer surplus at equilibrium fixed fees is given by:

$$CS^* = \frac{144tz\alpha + 180t^2 - \alpha^4 - 72tz^2 - 108t\alpha^2 + 72\Delta_I t\left(2\alpha - 2z - \Delta_E\right) - D\left(36t + D + 2\alpha^2\right)}{-144t}$$

Regarding consumers we have to ensure that all consumers have a positive surplus in any circumstance.

$$\begin{aligned}\frac{\left(z + \Delta_I\right)^2}{2} - tx^* - F_I^* &> 0 \Leftrightarrow\\ \left(\Delta_E\left(2z - 4\alpha + \Delta_E\right) - 4z\alpha - 6t + \Delta_I\left(2z + \Delta_I\right) + 2z^2 + 3\alpha^2\right) &> 0.\end{aligned}$$

This expression is minimized when $\Delta_E = \Delta_I = 0$ at $\left(2z^2 - 4z\alpha - 6t + 3\alpha^2\right) > 0$. Given our assumption (3) and $z > \alpha$ this is always verified.

Finally, total welfare is:

$$W^* = \frac{D\left(36t + 5D + 10\alpha^2\right) + 72t\left(z + \Delta_E\right)^2 + 5\alpha^4 - 36t\left(t + \alpha^2\right)}{144t}$$

For $\alpha > \sqrt{6t - D}$, the indifferent consumer is at $x > 1$, and therefore we do not have an interior solution. In this case, the optimal fixed fees and profits are:

$$\begin{aligned} \pi_I^* &= F_I^* = \frac{\left(z + \Delta_I\right)^2 - \left(z - \alpha + \Delta_E\right)^2}{2} - t \\ \pi_E^* &= F_E^* = 0, \end{aligned}$$

and consumer surplus and welfare are

$$\begin{aligned} CS^* &= \frac{\left(z - \alpha + \Delta_E\right)^2}{2} - \frac{3}{2}t \\ W^* &= \frac{\left(z + \Delta_I\right)^2}{2} - \frac{1}{2}t. \end{aligned}$$

This is true until we reach the monopoly fixed fee, i.e., for $\alpha \leq z + \Delta_E$. For $\alpha > z + \Delta_E$ the optimal fixed fee, profit, and welfare are as in the monopoly case. ■

Lemma 4: The second part is immediate from the observation of the entrant's profit function (11). With respect to the incumbent's profit function, we need to analyze it carefully. Taking the first and second derivatives we find that, for $\alpha_n < \sqrt{6t}$,

$$\begin{aligned} \frac{\partial \pi_I^{nn}\left(\alpha_n, C\right)}{\partial \alpha_n} &= \frac{1}{18t}\left(\alpha_n^3 - 30t\alpha_n + 18t\left(z + \Delta_d\right)\right) \\ \frac{\partial^2 \pi_I^{nn}\left(\alpha_n, C\right)}{\partial \alpha_n^2} &= \frac{1}{6t}\left(\alpha_n^2 - 10t\right) < 0. \end{aligned}$$

Moreover, we find that $\left.\frac{\partial \pi_I^{nn}(\alpha_n, C)}{\partial \alpha_n}\right|_{\alpha_n = \sqrt{6t}} = \left(z + \Delta_d\right) - \frac{4}{3}\sqrt{6t}$, which is positive given our assumption (3). Thus, $\pi_I^{nn}\left(\alpha_n, C\right)$ is increasing in α_n. For $\alpha_n \in \left[\sqrt{6t}, z + \Delta_d\right]$, we have

$$\frac{\partial \pi_I^{nn}\left(\alpha_n, C\right)}{\partial \alpha_n} = \left(z + \Delta_d\right) - \alpha_n > 0.$$

■

Lemma 6: The proof is similar to the previous one. ■

Lemma 7: For $\Delta_d = \overline{\Delta}_d$, we find that $\alpha_o^{\max} \equiv 0$, and thus, for higher values of Δ_d the entrant's market share will be zero for all $\alpha_o \geq 0$. ■

Lemma 8: Comparing $\pi_E^{nn}(\alpha_n)$ and $\pi_E^{no}(\alpha_o)$, for $\alpha_o < \alpha_o^{\min}$ and $\alpha_n < \sqrt{6t}$, we find that $\pi_E^{nn}(\alpha_n) > \pi_E^{oo}(\alpha_o)$ if and only if $\alpha_n < \sqrt{\alpha_o^2 + \Delta_d(2z + \Delta_d)}$. Thus, for $\alpha_n \geq \min\left\{\sqrt{6t}, \sqrt{\alpha_o^2 + \Delta_d(2z + \Delta_d)}\right\}$, the entrant will not ask for access to the NGN. The rest follows from Lemmas 3 and 5. ■

Lemma 9: Follows from Lemma 3. ■

Lemma 10: Taking the first derivative of $W^{nn}(\alpha_n, C)$, we obtain as candidates to extrema $\alpha_n = -\frac{3}{5}\sqrt{10t}$, $\alpha_n = 0$ and $\alpha_n = \frac{3}{5}\sqrt{10t}$. Taking the second derivative, we find that it is equal to 1 for $\alpha_n = \left|\frac{3}{5}\sqrt{10t}\right|$ and it is negative for $\alpha_n = 0$. Therefore, the candidate to maximizer is $\alpha_n = 0$, at which welfare is equal to $W^{nn}(0, C) = \frac{z^2}{2} - \frac{t}{4} - C$. Moreover, given that the minimizer $\frac{3}{5}\sqrt{10t}$ occurs before $\sqrt{6t}$, we need to check if it is not better to have a monopoly instead. For $\alpha_n \geq \sqrt{6t}$, welfare is given by $W^{Mn}(C) = \frac{z^2 - t}{2} - C < \frac{z^2 - 0.5t}{2} - C$, and therefore welfare is maximized at $\alpha_n = 0$. Finally, we find that $W^{nn}(0, C) > W^{no}(\alpha_o, C)$ for all $\alpha_o \geq 0$, which means that the regulator does not prefer to promote a duopoly on the old and new network. ■

Lemma 11: If the incumbent invests, and given Lemma 10, his ex-post profit will be $\pi_I^{nn}(0, C) = \frac{1}{2}t - C$. If he does not invest and $\alpha_o < \sqrt{6t}$, his profit is $\pi_I^{oo}(\alpha_o) = \frac{1}{2}t + \frac{\alpha_o^4 + 72tz\alpha_o - 60t\alpha_o^2}{72t}$, while if $\alpha_o \geq \sqrt{6t}$ it is $\pi^{Mn} = \frac{1}{2}z^2 - t$. Comparing, we find that $\pi_I^{nn}(0, C) < \pi_I^{oo}(\alpha_o)$ for all $\alpha_o < \sqrt{6t}$ and $\pi_I^{nn}(0, C) < \pi^{Mn}$ for all $\alpha_o \geq \sqrt{6t}$. ■

Lemma 12: See the proof of Lemma 10. ■

Proposition 15: From the results of Lemmas 2 to 12. ■

Lemma 14: Omitted. ■

Lemma 15: Taking the first derivative of $W^{no}(\alpha_o, C)$, we obtain as candidates to extrema $\alpha_o = \pm\sqrt{\frac{18}{5}t - 2z\Delta_d - \Delta_d^2}$ and $\alpha_o = 0$. Taking the second derivative, we find

that it is positive for $\alpha_o = \left|\sqrt{\frac{18}{5}t - 2z\Delta_d - \Delta_d^2}\right|$ and negative for $\alpha_o = 0$. Therefore, and given that $\sqrt{\frac{18}{5}t - 2z\Delta_d - \Delta_d^2} < \alpha_o^{\max}$, the candidates to maximizers are $\alpha_o = 0$ and $\alpha_o = \alpha_o^{\max}$.

According to the previous section, in order to have a duopoly on the old and new network, we need $\alpha_o \geq \alpha_o^{inv}(C)$ and $\alpha_o < \alpha_o^{\max}$. Whenever, for a given C, we have $\alpha_o^{inv}(C) \geq \alpha_o^{\max}$, it will be impossible to induce a duopoly on the old and new network. Note that for $\Delta_d \geq \overline{\Delta}_d$ this happens for all C.

Now assume that C is such that $\alpha_o^{inv}(C) < \alpha_o^{\max}$. In this case the regulator should either set $\alpha_o = \alpha_o^{inv}(C)$, which is the lowest access price which induces investment, or $\alpha_o = \alpha_o^{\max}$. This is due to the welfare function format analyzed above. If it is better, from a welfare point of view, to set $\alpha_o = \alpha_o^{\max}$, then a duopoly on the old and new network cannot optimal. If, on the other hand, it is socially better to set $\alpha_o = \alpha_o^{inv}(C)$, one can show that this solution is dominated by a duopoly on the new network. Indeed, if the regulator sets $\alpha_n = \widetilde{\alpha}_n\left(\alpha_o^{inv}(C)\right) = \sqrt{\alpha_o^{inv}(C)^2 + \Delta_d(2z + \Delta_d)}$, then $\pi_E^{nn}(\alpha_n) = \pi_E^{no}(\alpha_o)$, and $W^{nn}(\alpha_n, C) = W^{no}(\alpha_o, C)$. Moreover, we can show that $\pi_I^{nn}\left(\widetilde{\alpha}_n\left(\alpha_o^{inv}(C)\right), C\right) > \pi_I^{no}\left(\alpha_o^{inv}(C), C\right) = \pi_I^{oo}\left(\alpha_o^{inv}(C), C\right)$. Note that:

$$\begin{aligned} f(\alpha_o, C) &= \pi_I^{nn}\left(\widetilde{\alpha}_n(\alpha_o), C\right) - \pi_I^{no}(\alpha_o, C) \\ &= -\left(z\alpha_0 + 2z\Delta_d + \Delta_d^2 - \left(\sqrt{2z\Delta_d + \alpha_0^2 + \Delta_d^2}\right)(z + \Delta_d)\right) \end{aligned}$$

Function $f(\alpha_o)$ is decreasing in α_o because $\left.\frac{\partial f(\alpha_o)}{\partial \alpha_o}\right|_{\Delta_d = 0} = 0$ and $\frac{\partial^2 f(\alpha_o)}{\partial \alpha_o \partial \Delta_d} < 0$. Additionally, $f(z) = 0$. Hence, for all $\alpha_o < z$ we have that $f(\alpha_o) > 0$.

Therefore, if the regulator sets $\alpha_n = \widetilde{\alpha}_n\left(\alpha_o^{inv}(C)\right) - \varepsilon$, the incumbent still invests and the entrant asks for access to the new network, and $W^{nn}(\alpha_n, C)$ is higher than $W^{no}\left(\alpha_o^{inv}(C), C\right)$. ■

Lemma 16: First note that, given its definition, $\widehat{C}_2$ is such that for $C > \widehat{C}_2$, $\alpha_n^{inv}(\alpha_o, C)$ does not exist or is above $\alpha_n^{\max}(\alpha_o)$ for all α_o, which means that a duopoly on the new network is impossible for any α_n and α_o.

In order to reach the investment sub-game and to promote a duopoly on the new network, the regulator must set access prices such that $\alpha_n \geq \alpha_n^{inv}(\alpha_o, C)$. Given the format of the welfare function analyze previously, $\alpha_n = \alpha_n^{inv}(\alpha_o, C)$ is a candidate to maximizer for $C \leq \widetilde{C}(\alpha_o)$. Hence, and given that $\frac{\partial \alpha_n^{inv}(\alpha_o, C)}{\partial \alpha_o} > 0$, the regulator must set $\alpha_o = \widetilde{\alpha}_o(C)$, which is the lowest α_o such that $\alpha_n^{inv}(\alpha_o, C)$ is well defined, so that it is

possible to set the lowest $\alpha_n = \alpha_n^{inv}\left(\widetilde{\alpha}_o\left(C\right), C\right)$. We now just need to compare welfare at $\alpha_n = \alpha_n^{inv}\left(\widetilde{\alpha}_o\left(C\right), C\right)$ with monopoly welfare.

At $C = 0$, $\widetilde{\alpha}_o\left(C\right) = 0$, and thus $W^{nn}\left(\alpha_n^{inv}\left(\widetilde{\alpha}_o\left(0\right), 0\right), 0\right) = W\left(0, 0\right) > W^{Mn}\left(0\right)$. When C increases, $\widetilde{\alpha}_o\left(C\right)$ remains constant until $\widetilde{C}\left(0\right)$, and then, if $\widetilde{C}\left(0\right) < \widehat{C}_2$, $\widetilde{\alpha}_o\left(C\right)$ increases until we reach $\widehat{C}_2$, after which it is no more possible to have a duopoly on the new network. This happens because for $C \in \left[\widetilde{C}\left(0\right), \widehat{C}_2\right]$, $\widetilde{\alpha}_o\left(C\right)$ will be defined as the lowest α_o such that $\alpha_n^{inv}\left(\alpha_o, C\right) = \alpha_n^{\max}\left(\alpha_o, C\right)$, and thus, at $\widetilde{\alpha}_o\left(C\right)$, $\alpha_n^{inv}\left(\alpha_o, C\right)$ must be cutting $\alpha_n^{\max}\left(\alpha_o\right)$ from above. In this case, when C increases, the two cannot intercept to the left of the initial point, which implies that $\widetilde{\alpha}_o\left(C\right)$ is not decreasing. From this, one can conclude that for $C \in \left[\widetilde{C}\left(0\right), \widehat{C}_2\right]$, $W^{nn}\left(\alpha_n^{inv}\left(\widetilde{\alpha}_o\left(C\right), C\right), C\right) - W^{Mn}\left(C\right)$ is decreasing in C. There may then be a $\widehat{C}_1 < \widehat{C}_2$, such that $W^{nn}\left(\alpha_n^{inv}\left(\widetilde{\alpha}_o\left(\widehat{C}_1\right), \widehat{C}_1\right), \widehat{C}_1\right) = W^{Mn}\left(\widehat{C}_1\right)$. If there is no such $\widehat{C}_1$, the regulator starts inducing to an investment sub-game through monopoly for $C > \widehat{C}_2$. If $\widetilde{C}\left(0\right) = \widehat{C}_2$, there is no $\widehat{C}_1$, and the previous statement is also valid. ■

Proposition 17: If $C \geq \frac{\Delta_d(2z+\Delta_d)}{2}$, social welfare is higher at the no-investment sub-game. In fact, at the no-investment sub-game the maximum welfare is equal to $W^{oo}\left(0\right) = \frac{z^2}{2} - \frac{1}{4}t$, while at the investment sub-game the maximum welfare is strictly lower than $W^{nn}\left(0, C\right) = \frac{\left(z+\Delta_d\right)^2}{2} - \frac{1}{4}t - C$ since $\alpha_n > 0$. Comparing both, we find that $W^{oo}\left(0\right) \geq W^{nn}\left(0, C\right)$ for $C > \frac{\Delta_d(2z+\Delta_d)}{2}$.

If $C \leq \frac{\Delta_d(2z+\Delta_d)}{2} - \frac{1}{4}t$, social welfare is higher at the investment sub-game. In fact, at the investment sub-game the regulator has at least assured the monopoly welfare level (see case v) of the previous section). Comparing this with the no-investment sub-game welfare we find that $W^{Mn}\left(C\right) \geq W^{oo}\left(0\right)$ for $C \leq \frac{\Delta_d(2z+\Delta_d)}{2} - \frac{1}{4}t$.

As the distortions introduced on the access prices which maximize welfare in the investment sub-game are increasing on C there will be a $\overline{C} \in \left(\frac{\Delta_d(2z+\Delta_d)}{2} - \frac{1}{4}t, \frac{\Delta_d(2z+\Delta_d)}{2}\right)$ above which it is socially optimal better to induce to a no investment sub-game equilibrium. ■

Proposition 18: From the previous Lemmas. ■

Proposition 19 : Just compare (25) and (26). ■

Corollary 20: Omitted. ■

Corollary 21: Omitted. ■

Proposition 22: The regulator maximizes $W(\alpha_n, C)$ subject to the restrictions: $\pi_I^{nn}(\alpha_n, C) + P_n \geq \pi_I^{oo}(\alpha_o) + P_o$ and $\pi_E^{nn}(\alpha_n) - P_n > 0$. In order to achieve the first best these cannot be binding at $\alpha_n = 0$, which is verified when $\frac{1}{2}t - C + P_n \geq \pi_I^{oo}(\alpha_o) + P_0$ and $\frac{1}{2}t > P_n$. The RHS of the first restriction is minimized at $\alpha_o = 0$ and $P_o = 0$, where it becomes $\frac{1}{2}t - C + P_n \geq \frac{1}{2}t$, or equivalently $P_n \geq C$. Joining the two restrictions, we find that they are not binding when $\frac{1}{2}t > C$. ■

Proposition 23: The prove is similar to the previous Proposition, with the difference that the second restriction is now $\pi_E^{nn}(0) - P_n > \max\{\pi_E^{no}(\alpha_o) - P_o, 0\}$.

Assume that $\pi_E^{no}(\alpha_o) \geq P_o$. The two restrictions can then be summarized by $P_n - P_o \in [\pi_I^{oo}(\alpha_o) - \pi_I^{nn}(\alpha_n, C), \pi_E^{nn}(\alpha_n) - \pi_E^{no}(\alpha_o))$. These will not be binding at $\alpha_n = 0$ if $\pi_I^{oo}(\alpha_o) - \pi_I^{nn}(\alpha_n, C) < \pi_E^{nn}(\alpha_n) - \pi_E^{no}(\alpha_o)$, or equivalently $t > C + \pi_I^{oo}(\alpha_o) + \pi_E^{no}(\alpha_o)$.

Assume now that $P_o \in [\pi_E^{no}(\alpha_o), \pi_E^{oo}(\alpha_0))$. At $\alpha_n = 0$, restrictions become $\frac{1}{2}t > P_n$ and $P_n - P_o \geq \pi_I^{oo}(\alpha_o) - \frac{1}{2}t + C$. The LHS of the latter is maximized if we choose the lowest P_o and the highest P_n possible, where it becomes $\frac{1}{2}t - \pi_E^{no}(\alpha_o) > \pi_I^{oo}(\alpha_o) - \frac{1}{2}t + C$, or $t > C + \pi_I^{oo}(\alpha_o) + \pi_E^{no}(\alpha_o)$.

Finally, for the case where $P_o \geq \pi_E^{oo}(\alpha_0)$, restrictions become $\frac{1}{2}t > P_n$ and $P_n \geq \pi_I^{Mo}(\alpha_o) - \frac{1}{2}t + C$, which are not binding if $\frac{1}{2}t > \pi_I^{Mo}(\alpha_o) - \frac{1}{2}t + C$ or equivalently $t > \pi_I^{Mo}(\alpha_o) + C$. This condition is stronger than the previous cases condition, and therefore, the regulator should not use such P_o. ■

References

Biglaiser, G., and DeGraba, P., 2001, "Downstream integration by a bottleneck input supplier whose regulated wholesale prices are above costs", *Rand Journal of Economics*, 32, 2, 302-315.

Foros, O., 2004, "Strategic Investments with Spillovers, Vertical Integration and Foreclosure in the Broadband Access Market", *International Journal of Industrial Organization,* 22, 1-24.

Guthrie, G., 2006, "Regulating Infrastructure: The Impact on Risk and Investment", *Journal of Economic Literature*, 44, 925-972.

KOTAKORPI, K., 2006, "Access Price Regulation, Investment and Entry in Telecommunications", *International Journal of Industrial Organization,* 24, 1013-1020.

VALLETTI, T., 2003, "The Theory of Access Pricing and its Linkage with Investment Incentives", *Telecommunications Policy,* 27, 659-675.

VAREDA, J., and HOERNIG, S., 2007, "The race for telecoms infrastructure investment with bypass: Can access regulation achieve the first best?", CEPR Discussion Paper 6012.

VAREDA, J., 2007, "Unbundling and the Incumbent Investment in Quality Upgrades and Cost Reduction", FEUNL working paper 526.

Vertical separation and network investment in telecommunications

Alessandro Avenali, Giorgio Matteucci, Pierfrancesco Reverberi

Dipartimento di Informatica e Sistemistica "Antonio Ruberti"
Sapienza – Università di Roma
Via Ariosto, 25 – 00185 Roma, Italia
e-mail: {avenali, matteucci, reverberi}@dis.uniroma1.it

Abstract. When the access network is an enduring economic bottleneck, vertical separation of the telecommunications incumbent may be an effective and proportionate remedy. There is the presumption that separation would reduce quality-enhancing network investment. We show that, despite efficiency losses of vertical disintegration, mandatory separation improves quality investment and welfare provided that the demand-side investment spillover, or the rival firm's (perceived) service quality is sufficiently high. We find that separation mostly encourages investment when the integrated firm provides downstream competitors with low-quality access. The results shed light on the effect of separation on the incentive to deploy Next Generation Access Networks.

1 Introduction

The situation where a dominant firm controls the supply of an essential input, while there is at least potential competition in both downstream and upstream markets, is common to network industries such as electricity, gas, railways, and fixed telecommunications, and so is the issue of preventing the dominant firm from leveraging her market power into vertically related markets[1]. Notwithstanding these common features, network industries in Europe are characterized by different institutional settings. On the one hand, in energy sectors as well as railways regulators and governments have often designed the industry structure by imposing various forms of vertical separation of the bottleneck input owner. On the other hand, in fixed telecommunications the dominant undertaking has usually been controlled by the exclusive use of behavioural remedies, such as access obligations, while preserving vertical integration of networks and services[2].

[1] The bottleneck input respectively is electricity transmission, gas transportation, railway track, and access to the fixed telecommunications network (the so-called local loop).
[2] However, accounting separation has generally been imposed on the incumbent firm to improve effectiveness of behavioural remedies.

In the recent past, an extensive academic and policy debate concluded that the benefits of separation in fixed telecommunications are uncertain while the costs potentially large (OECD, 2003). In fact, mandatory separation of the local access owner would threaten the loss, or reduction of efficiencies enjoyed by an integrated firm, including economies of scale and scope, and would provoke an increase in transaction costs. Thus, the impact on consumers in terms of reduced prices and improved service quality would be uncertain. Furthermore, there would be a considerable one-off cost of divestment. Above all, it was not certain whether or not local access would be remained an enduring economic bottleneck (de Bijl, 2004a), so that the key question underlying the case for vertical separation was left unanswered.

Empirical evidence in the last years has shed some light on these issues. First, it has become clear that there is not any effective wide-scale competitive technology to the traditional copper lines of incumbent firms. To date, many alternative access infrastructure deployments (examples include cable and fibre networks) have been focussed on specific geographic locations and lack the economies of scale that national deployments may benefit from. Furthermore, wireless technologies such as mobile or Wi Max are not able to provide a complete set of effective substitutes for end-users in broadband access markets, especially in high-density urban areas.

Second, it has also become clear that behavioural regulation aimed at preventing both price and non-price discrimination on the part of the incumbent foreclosing downstream markets is often highly intrusive or ineffectual. Indeed, regulation is most effective when defines remedies which the dominant firm has the incentive to adhere to. Thus, in exceptional circumstances, or following repeated breaches of contract, the regulator should be able to treat structural competition problems by structural (rather than behavioural) remedies. Given that wireline access networks can be considered as an enduring economic bottleneck, the extreme regulatory option should include vertical separation of the integrated firm.

Vertical separation allows the regulator to focus intervention on those market segments where replication of the incumbent's assets is infeasible or economically undesirable. Furthermore, it creates an environment where the access owner has the incentive to behave consistent with an equal-access development of competition[3]. Faulhaber (2003) emphasizes

[3] OFCOM (the UK regulator) has recently obtained that British Telecom (BT) adheres to functional separation and provides some essential wholesale services to downstream competitors on an 'equivalence of inputs' basis (OFCOM, 2005). This means that BT has committed to behavioural and organizational changes to separate bottleneck activities from competitive ones and ensure that rival firms benefit from access truly equivalent to the incumbent's own retail businesses. Successively, the European Regulators Group (ERG) has claimed that functional separation reinforces the existing remedies and complements them by ensuring that National Regulatory Authorities (NRAs) can intervene where particularly non-discrimination behaviour cannot be

that the vertical disintegration of the Bell System in the US has been crucial to achieve the goal of promoting fair competition, since it has ensured that the local access owner has the private incentive to provide equal access to any downstream firm.

However, a number of questions remain still open. One of the most critical issues related to mandatory separation of the integrated firm is the adverse effect this may produce on investments in network quality. Such a negative effect would follow from the fact that vertical separation deprives the dominant firm of retail revenues, and may impede the desirable coordination of upstream and downstream investment activities. This is a highly relevant issue in fixed telecommunications, particularly in the current phase where in several countries investment in next generation access networks (NGANs) continues to gather pace, with announced deployments mainly from incumbent firms, but also from some Other Licensed Operators (OLOs)[4]. Compared with current generation wireline access networks, NGANs can be used to deliver significantly higher bandwidths, which support the development of new value-added interactive services (such as High-Definition IP-TV)[5]. The availability of a higher speed access infrastructure would also produce benefits that accrue to the economy of a country as a whole through increased productivity and competitiveness.

The unique characteristics of NGANs are likely to have a profound effect on competition and industry structure, thus posing new regulatory challenges. On the one hand, the prospective deployment of NGANs is based on technology developments that raise OLOs' investment costs, while limiting the possibilities for co-location and undermining the progress recently achieved in intra-platform competition, particularly through local loop unbundling[6].

---

addressed through other remedies (ERG, 2007). Thus, in the very recent review of the regulatory framework (NRF) of electronic communications markets, the European Commission has included functional separation within the set of remedies that NRAs have at their disposal to promote competition in relevant markets.

[4] NGANs are realized by extending the fibre network closer to the customer premises, either to the home (FTTH) or, more frequently, to the street cabinet (FTTC). NGAN investment plans include Deutsche Telekom's, KPN's and Telecom Italia's FTTC deployments respectively in Germany, the Netherlands and Italy, Illiad's proposed FTTH deployment in French metropolitan areas, AT&T and Verizon's investments in higher speed broadband access networks in the US, and NTT's deployment of FTTH in Japan.

[5] Capacity constraints may mean that wireless networks are less suitable for such high-bandwidth applications.

[6] Empirical evidence shows that competitive access network roll out in the EU has been targeted at business customers or urban areas, so that there is very limited replication of the incumbent's access network. Intra-platform competition is largely predominant over inter-platform competition, and xDSL (80.3% of retail broadband lines) cannibalize other technologies, among which cable modem prevails (COCOM, 2007). Currently, infrastructured OLOs operate via local loop unbundling (LLU), which requires them to incur the costs related to installing their network equipment (such as DSLAMs) at local exchanges and to connecting their backhaul network to co-location sites. OLOs can thus upgrade the access network from narrowband to broadband use and provide differentiated services from those of the incumbent. A number of studies estimate that the strong local economies of scale effects that are evident in fibre deployment at the cabinet, combined with the restricted scope for service differentiation and the uncertainty about consumers' willingness to pay for new services, actually mean that the use of sub-loop unbundling (SLU), which requires OLOs to further deploy their networks, is not economically viable as an alternative to LLU to reach the mass market (see e.g. Analysis, 2007).

Consequently, in those countries where there is insufficient inter-platform competition, mandatory separation of the local access owner (with suitably regulated charges to new access infrastructures) may be essential to prevent market foreclosure as a result of technology developments creating or reinforcing enduring economic bottlenecks, in the absence of appropriate regulation.

On the other hand, the deployment of NGANs imposes on bottleneck asset owners both significant costs (which, notwithstanding the advantages of incumbency, are not completely sunk) and risks (related to demand uncertainty inherent in new services). Consequently, incumbents have been arguing that a regulatory exemption regime is vital to maintain the economic rationale to roll out new access infrastructures. In fact, they claim that there is little point to ex ante regulation that concerns itself with the potential for abusive dominance to arise from the build-out of NGANs if the threat of such regulation delays or even prevents these networks from being built at all, thus being detrimental to social welfare.

The purpose of this paper is to study how the vertical structure of the industry affects the local access owner's incentives to undertake network investments improving service quality, and more generally social welfare. For this purpose, we compare both quality investment and welfare under the two alternative scenarios where the access network owner respectively is vertically separated or integrated into the downstream broadband market. Vertical separation here implies splitting the incumbent into a company managing the bottleneck activities, that is, owning the local access network and providing wholesale access (so-called Loop Co), and a company managing competitive activities, that is, owning the long-distance backbone and providing retail services. Thus, the latter company buys access and leases local lines from Loop Co, just like resale-based and LLU-based (or SLU-based) OLOs. For the sake of clarity, in what follows we focus on structural separation, where Loop Co achieves profit entirely in the wholesale market and takes investment decisions on the sole basis of that profit[7].

We assume imperfect downstream price competition with differentiated products and partial market participation, which can be considered as two basic features of the retail broadband mass market. We assume that, both under vertical integration and separation, the incumbent has a higher reputation than the rival firm (or, equivalently, there are consumer

[7] However, the model can be easily adapted to deal with the case of functional separation, where the upstream business unit has the obligation to treat both affiliated and unaffiliated downstream units in a perfectly equivalent manner, but may still take a strategic decision such as investing in network quality by considering the whole company's profit (i.e. the sum of the affiliated units' upstream and downstream profits), while avoiding unfair cross-subsidies.

switching costs). Moreover, under vertical integration the incumbent has a higher ability than the rival firm to offer value-added services that benefit from investment in network quality[8].

In this framework, we find that vertical separation may raise both quality investment and welfare compared with vertical integration, albeit we consider efficiency losses due to the incumbent's disintegration. These results particularly occur when either the downstream firms' ability to offer advanced services, or the rival firm's (perceived) service quality is high enough. In each of these cases, inducing a level-playing field competition by means of vertical separation produces the highest social benefits. In contrast with the prevailing thesis in the relevant literature, we argue that there is not an evident trade-off between inducing competition through vertical separation and ensuring the access owner's network investment[9].

This paper is organized as follows. Section 2 discusses the literature. Section 3 first presents the model, and then analyzes and compares the two alternative scenarios respectively of vertical integration and separation. Section 4 contains some concluding remarks.

2 Relevant literature

While there has been wide research on principles for pricing access to a vertically integrated firm's network in a static environment (Armstrong, 2002), there is still a limited literature on the dynamic properties of access pricing rules, and on their effect on incentives to invest. We focus here on network quality investments in a one-way access setting, that is, an asymmetric setting where the rival firm has not yet fully developed his network[10].

Promoting competition and ensuring the access network owner's investment are perceived as conflicting goals, so that light-handed regulation is deemed essential to encourage investment (see e.g. Pindyck, 2007). Even a temporary regulatory forbearance (the so-called access holiday) is invoked to preserve the incumbent's incentive to undertake risky

[8] Thus, under integration the demand-side spillover of quality investment is higher to the facility-based firm's retail subsidiary than to the independent firm. Such a higher ability to transform input into output may either follow from premium content provision or from non-price discrimination degrading the input quality provided to the downstream competitor.

[9] Although quantitative results differ, these qualitative remarks hold independent of the form of separation. Indeed, when closely monitored functional separation may effectively promote competition, while preserving some efficiencies accruing from integration. See Cave (2006) for more details on the different forms of separation, and their likely effect on competition. See also OECD (2006) for a review on the pro-competitive effects of vertical separation in a number of countries.

[10] We do not consider the linkage between access pricing and cost-reducing investment. We also do not review the literature on two-way access pricing and investment incentives (see e.g. Valletti and Cambini, 2005). Recent work in this field considers using asymmetric regulation to foster OLOs' investments (Peitz, 2005).

infrastructure investment (Gans and King, 2004). Some recent models include two basic features that we also embrace here, that is, partial market participation and imperfect downstream competition. In this framework, Foros (2004) finds that access price regulation reduces quality investment, and may also reduce consumer surplus.

Some papers consider the effect of multi-period access pricing schedules on the OLOs' incentive to build alternative networks. Bourreau and Dogan (2005, 2006) show that the availability of a resale product (both at a constant and at a time-varying price) retards the OLO's investment in a new technology. They argue that the incumbent that initially resists mandated access is aware that her network becomes less essential over time. Hence, to avoid facility-based competition the incumbent voluntarily charges attractive access prices. The policy implication entails banning resale entry when infrastructure competition becomes both feasible and socially desirable, while setting a sunset clause on regulation is not essential[11].

The main finding is based on the assumption that the cost of technology adoption declines over time, so that facility-based entry ultimately takes place whether or not service-based entry has previously occurred. Some authors take the opposite view that the less replicable assets should have a low access price to induce entrants to deploy complementary systems (Cave and Vogelsang, 2003; de Bijl and Peitz, 2004b). The idea that service-based and facility-based competition are complement and not substitute entry modes lies at the heart of the ladder of investment paradigm, which is currently the prevailing regulatory model in the EU (ERG, 2006). According to this paradigm, developing an alternative network is not so much a question of time *per se* as is related to building a customer base that increases reputation and brand loyalty to the OLO, and reduces the (unit) cost and the risk of network investment. By changing the incentive properties of access regulation over time, the NRA can induce OLOs to gradually roll out their networks.

While the above cited papers consider behavioral (access) regulation of a vertically integrated firm, there are also a few papers that analyze formally how different institutional settings affect incentives to invest in network quality. Buehler et al. (2004) assume that quality is difficult and costly to specify, and therefore cannot be described *ex ante* in a regulatory contract and ascertained *ex post* by a court (this is the same as in our paper, where quality is observable but non-verifiable). They find that, in the great majority of cases, the network owner's quality investment is smaller under vertical separation than integration[12].

[11] Sunset clauses specify a period of time after which access to the incumbent's network is no longer regulated.
[12] Nonetheless, quality investment incentives under separation are the same as under integration when the NRA is allowed to use a suitable two-part tariff for charging network access.

However, their model considers a chain of monopolies (possibly with competition *for* the retail market among a number of identical firms, in which case the downstream monopoly is auctioned off). Thus, the authors do not take account of the case which is most relevant to the retail broadband market, that is, imperfect price competition with differentiated products.

Buehler et al. (2006) extend previous research by allowing for varying degrees of retail competition, but downstream firms sell a homogeneous product and compete on quantities. They find that NRAs face a price vs. quality trade-off when opening up a vertically integrated monopoly to downstream competition (and possibly banning the incumbent firm from the retail market), as this cannot yield both a lower retail price and higher network quality. They also find that there is yet scope for welfare-improving competition by suitably trading off network quality against lower consumer prices (even if tough competition is necessary to achieve this goal under vertical separation). However, Buehler et al. (2006) do not assess quality investment incentives under vertical integration (with downstream competition) directly against vertical separation. While comparing these vertical structures might not be essential in such network industries as energy and railways, it turns out to be critical as far as fixed telecommunications is concerned.

In this paper, we compare both quality investment and welfare when the facility-based firm respectively is vertically separated or integrated into the retail broadband market, where there is price competition under product differentiation. Since we suitably consider the role of downstream competition in promoting network investment, then in a number of circumstances we are able to reverse previous findings and obtain that the network owner's investment is higher under vertical separation than integration (see Section 3.3).

3 The model

We consider two alternative scenarios. In the first one, a vertically integrated firm (firm *i*) competes in prices with a rival firm (firm *e*) in the retail broadband access market. We assume that the local access network is not duplicable, so that firm *e* needs to buy the essential input from the integrated firm in order to provide his retail service[13]. We also assume that the wholesale access charge is set by the regulator, while the retail market is not regulated[14].

[13] To prevent confusion, from now on we refer to firm *i* as "she" and firm *e* as "he".
[14] The newly revised NRF prescribes that regulation focus on bottlenecks and retail remedies be withdrawn as far as possible. While wholesale broadband access is in the list of relevant markets, retail broadband access is not.

The vertically integrated firm undertakes infrastructure investment to upgrade the quality of the access network, such that both firm *i*'s retail subsidiary and (due to the spillover effect) the rival firm benefit from an increase in consumer willingness to pay (henceforth, wtp for brevity) for the value-added services they are enabled to provide on the basis of the network investment[15]. We assume that retail services are vertically differentiated and consumers have a higher wtp for the incumbent's service, both because there are first-mover advantages (such as brand loyalty, reputation, or consumer switching costs), and firm *i*'s retail subsidiary has a higher ability than the rival firm to benefit from the upstream quality investment[16].

In the second scenario, we assume that the entity that manages the local network (firm *a*)[17] is vertically separated, so that it provides wholesale access and undertakes quality investment, but does not have any affiliated entity competing in the downstream market. In such a case, the upstream firm derives its profit solely from selling wholesale access to downstream firms (at a regulated access charge).

We assume that brand loyalty or switching costs still favor the former retail subsidiary of the integrated firm. However, under vertical separation both downstream firms are provided with the same input quality at the same access charge. Thus, we assume that under separation downstream firms have the same ability to transform input into output, that is, they equally benefit from network quality investment. Vertical separation is also the source of some inefficiencies[18]. We assume that the loss of economies of scope and the lack of coordination between the wholesale and retail markets reduce the downstream firms' ability to benefit from quality investment compared with the integrated firm's retail subsidiary.

We define a three-stage game of complete information. The timing is as follows:

1. The regulator sets the network access charge w.
2. The local access owner (either vertically integrated or separated) sets the level of investment in network quality x.
3. The incumbent firm (either vertically integrated or separated) and the rival firm simultaneously choose retail prices p_i and p_e.

[15] For simplicity, we assume that there is no uncertainty about costs and returns on network quality investment.
[16] The source of the latter competitive advantage can be twofold. On the one hand, firm *i* may provide a larger variety of services than firm *e*, or may have exclusive or privileged access to premium content compared with the rival firm with a smaller customer base. On the other hand, as a possible response to access price regulation firm *i* may provide the downstream competitor with a lower-quality input than her subsidiary, according to a non-price discrimination strategy (see e.g. Mandy and Sappington, 2007).
[17] From now on, we refer to firm *a* as "it".
[18] Since the local access owner's profit stems entirely from the upstream market, then the (regulated) wholesale access charge may be expected to rise compared with vertical integration.

In the following sections, we solve the model both under vertical integration and separation, and compare the results obtained to evaluate how vertical separation affects incentives to invest in network quality and, more generally, social welfare.

3.1 Vertical integration

Let $s+x$ be consumer *s*'s valuation of the incumbent's product, where s is the consumer's wtp for the basic service (that is, broadband internet access), which is uniformly distributed within the interval $[0,1]$, and x is the increase in wtp for the value-added services that firm *i* may offer on the basis of the quality-improving investment in the access network. On the other hand, consumer *s*'s valuation of the rival firm's product is $\gamma \cdot s + \delta \cdot x$, where $(1-\gamma)$ is the consumer switching cost, while δ measures the spillover effect, that is, firm *e*'s ability to transform one unit of quality investment into services that are valuable to end-users[19]. For simplicity, we assume that $\gamma \in (2/3,1)$ and $\delta \in (2/3,1)$.

We assume that consumers have unit demands. If $s + x - p_i > \gamma \cdot s + \delta \cdot x - p_e$ then consumer s decides to buy from the incumbent rather than the rival firm, because of a higher net utility (otherwise, consumer s buys from firm *e*). However, if net utilities are both negative, then consumer s neither buys from the incumbent nor from the rival firm. Thus, we allow for partial market participation, where low-wtp consumers may not be active.

We derive the demand curves of the two firms by identifying the locations of two specific consumers. The first consumer, denoted as s_{ind}, is the one that is indifferent between buying from either of the two firms. It follows that $s_{ind} + x - p_i = \gamma \cdot s_{ind} + \delta \cdot x - p_e$ must hold (where net utilities are both positive). Hence, we have $s_{ind} = \dfrac{p_i - p_e - (1-\delta) \cdot x}{1-\gamma}$. The second consumer, denoted as s_{mar}, is the marginal consumer, that is the one that is indifferent between buying from the entrant or not buying at all. It follows that $\gamma \cdot s_{mar} + \delta \cdot x - p_e = 0$ must hold. Hence, we have $s_{mar} = \dfrac{p_e - \delta \cdot x}{\gamma}$. Since the (perceived) quality of the incumbent's product is higher than the rival firm's product, then high-wtp consumers buy from firm *i*. Since consumers have unit demands and are uniformly distributed within the interval $[0,1]$ then firms' demand curves are linear, and can be expressed respectively as $q_i = 1 - s_{ind}$ and

[19] Alternatively, $1-\delta$ measures the reduction in the quality of the input provided by the integrated firm to the rival firm compared with the input quality she provides to her retail subsidiary.

$q_e = s_{ind} - s_{mar}$, where $1 \geq s_{ind} \geq s_{mar} \geq 0$ must hold for satisfying feasibility constraints on quantities (i.e. $q_i \geq 0$, $q_e \geq 0$, and $q_i + q_e \leq 1$). Inserting for s_{ind} and s_{mar}, we obtain that:

$$q_i = 1 + \frac{1}{1-\gamma}\left((1-\delta)\cdot x - p_i + p_e\right); \quad q_e = \frac{1}{1-\gamma}\left(\frac{(1-\delta)}{\gamma}\cdot x + p_i - \frac{1}{\gamma}p_e\right).$$

We assume that the integrated firm has an upstream constant marginal (per-user) cost of providing the essential input, which, without loss of generality, we normalize to zero. We also assume that firm *i* faces a quadratic cost $C(x)=x^2/2$ related to investment in network quality, which is for every potential user. For simplicity, we normalize to zero any downstream cost.

Let w be the wholesale charge that the downstream competitor pays for access to the integrated firm's local network. Thus, firms' profit functions can be written as:

$$\pi_i = p_i q_i + w q_e - \frac{x^2}{2}; \quad \pi_e = p_e q_e - w q_e.$$

Inserting for quantities, we have:

$$\pi_i = \left(1 + \frac{1}{1-\gamma}\left((1-\delta)\cdot x - p_i + p_e\right)\right)p_i + \frac{w}{1-\gamma}\left(\frac{(1-\delta)}{\gamma}\cdot x + p_i - \frac{1}{\gamma}p_e\right) - \frac{x^2}{2};$$

$$\pi_e = \frac{p_e - w}{1-\gamma}\left(\frac{(1-\delta)}{\gamma}\cdot x + p_i - \frac{1}{\gamma}p_e\right).$$

We define social welfare $W_I = \pi_i + \pi_e + CS$ as the sum of firms' profits and consumer surplus $CS = \int_{s_{ind}}^{1}\left(s + x - p_i\right)ds + \int_{s_{mar}}^{s_{ind}}\left(\gamma\cdot s + \delta\cdot x - p_e\right)ds$ under vertical integration.

Solving the game backwards, first we find the third-stage optimal retail prices by the first order condition on firms' profits (given that the second order condition is always fulfilled):

$$p_i = \frac{3w + 2(1-\gamma) + (2-\gamma-\delta)x}{4-\gamma}; \quad p_e = \frac{(2+\gamma)w - \gamma(x-1+\gamma) + \delta(2-\gamma)x}{4-\gamma}.$$

At the second stage, firm *i* maximizes her profit with respect to the investment level x. Depending on both the investment spillover to firm *e* and the wholesale access charge (set by the NRA at the first stage), quality investment might increase consumers' wtp in such a way that either $s_{mar}(x(w)) < 0$ or $s_{ind}(x(w)) < 0$. This means that, in the former case, even the consumer located at $s = 0$ purchases from firm *e* (since $s_{mar} \geq 0$ must hold). Thus, we obtain the optimal investment by solving the equation $s_{mar}(x(w)) = 0$ with respect to *x*. In the latter case, the consumer located at $s = 0$ purchases from firm *i* (which thus achieves a downstream monopoly), and we obtain the optimal investment level by solving $s_{ind}(x(w)) = 0$. In all of the

remaining cases (i.e. for intermediate values of both the investment spillover and the access charge), the optimal quality investment is consistent with a downstream duopoly where low-wtp consumers are not active, and is obtained by the first order condition on firm *i*'s profit. The following proposition formalizes the results.

Proposition 1. At the second stage of the game, depending on both the spillover effect and the wholesale access charge, the following outcomes are possible in terms of quality investment and industry structure:

i) if both $\gamma \le \delta \le \frac{1}{2}\left(3\gamma - \gamma^2\right)$ *and* $w' \le w \le w''$ *hold then firm i's optimal investment level is* $x_{int}(w) = \frac{(1-\gamma)\left((4-w)\gamma^2 - 4\gamma(2-\delta) + 8\delta w\right)}{\gamma\left(-7\gamma^2 + \gamma^3 + 4\gamma(4+\delta) - 2(4+(4-\delta)\delta)\right)}$, *and there is a downstream duopoly where the lowest-wtp consumers do not buy;*

ii) otherwise, we have that:

a) if $w \le w_{ind} = \frac{(1-\gamma)\delta}{1-\delta}$ *then the optimal investment level is* $x_{mar}(w) = \frac{\gamma - \gamma^2 + w(2+\gamma)^2}{\gamma + 2\delta}$, *and there is a downstream duopoly where the marginal consumer is located at* $s = 0$*;*

b) if $w > w_{ind}$ *then investment is* $x_{ind}(w) = \frac{(1-\gamma)(2+w-\gamma)}{2-\gamma-\delta}$, *and there is a downstream monopoly where the marginal consumer is located at* $s = 0$ *and buys from firm i.*

Proof. See Appendix, where we relegate the expressions of w' and w''.

At the first stage, the regulator maximizes welfare with respect to the access charge. We distinguish two cases. If firm *e*'s ability to exploit firm *i*'s investment is sufficiently high (that is, if $2/3 < \delta'(\gamma) \le \delta \le 1$), then market size enlarges to the extent that the marginal consumer is the one with the lowest wtp (that is located at $s = 0$), which purchases from firm *e*. Alternatively, if the investment spillover is limited (that is, if $2/3 < \delta < \delta'(\gamma)$) then quality investment provides firm *i* with a competitive advantage. In such a case, there is a downstream monopoly where the lowest-wtp consumer (located at $s = 0$) purchases from firm *i*[20]. Proposition 2 below formalizes the results, while the following Table 1 summarizes the outcome of the game in terms of firms' market shares, quality investment, access charge, and social welfare.

[20] In both cases, the net utility of the marginal consumer in *s*=0 is equal to zero.

Proposition 2. At the equilibrium of the game, there is a critical value of the spillover effect $\delta'(\gamma) = \frac{4+\gamma}{2} - \frac{\sqrt{3(4-\gamma^2)}}{2} \in (2/3, 1)$ *such that:*

i) if $1 > \delta \geq \delta'(\gamma)$ *then the regulator sets the optimal access charge* $w^i_{mar} = \frac{(\gamma-1)\left(\gamma(1+\gamma)^2 - (-3+\gamma)\delta + (1+2\gamma)\delta^2\right)}{3(-2+\delta)\delta - 1 + 3\gamma^2 + \gamma^3 + 2\gamma(-1+\delta)^2}$, *and there is a downstream duopoly where the marginal consumer is located at* $s = 0$ *;*

ii) if $\frac{2}{3} \leq \delta < \delta'(\gamma)$ *then the optimal access charge is* $w^i_{ind} = \frac{(1-\gamma)\delta}{1-\delta}$, *and there is a downstream monopoly where the marginal consumer located at* $s = 0$ *buys from firm i.*

Proof. See Appendix.

	$\frac{2}{3} < \delta < \delta'(\gamma)$	$\delta'(\gamma) \leq \delta < 1$
q_i	1	$\frac{\gamma^2(2+\gamma) - 3\gamma\delta + 2(1+\gamma)\delta^2 - 2(1+\delta)}{3\gamma^2 + \gamma^3 + 2\gamma(-1+\delta)^2 + 3(-2+\delta)\delta - 1}$
q_e	0	$\frac{(\gamma+1)^2 - (4+\gamma)\delta + \delta^2}{3\gamma^2 + \gamma^3 + 2\gamma(-1+\delta)^2 + 3(-2+\delta)\delta - 1}$
x	$x^i_{ind} = \frac{1-\gamma}{1-\delta}$	$x^i_{mar} = \frac{(\gamma-1)(3+\delta+\gamma(3+\gamma+\delta))}{3\gamma^2 + \gamma^3 + 2\gamma(-1+\delta)^2 + 3(-2+\delta)\delta - 1}$
w	$w^i_{ind} = \frac{(1-\gamma)\delta}{1-\delta}$	$w^i_{mar} = \frac{(\gamma-1)\left(\gamma(1+\gamma)^2 - (-3+\gamma)\delta + (1+2\gamma)\delta^2\right)}{3\gamma^2 + \gamma^3 + 2\gamma(-1+\delta)^2 + 3(-2+\delta)\delta - 1}$
W_I	$\frac{2 - \gamma^2 + 2\gamma\delta + (\delta-4)\delta}{(1-\delta)^2}$	$\frac{2\gamma^3 + 2(-3+\delta)\delta + \gamma^2(5+2\delta) + \gamma(1+3(-2+\delta)\delta) - 3}{2\left(-1+3\gamma^2+\gamma^3+2\gamma(-1+\delta)^2+3(-2+\delta)\delta\right)}$

Table 1. Vertical integration – outcome of the game.

3.2 Vertical separation

Let us now assume that the entity that manages the local network is vertically separated. Let $\beta \cdot x$ be the increase in consumers' wtp for the value-added services provided by any downstream firm on the basis of the upstream firm's network quality investment, where

$1 > \beta > \delta$ [21]. Consumer s purchases the incumbent's product if $s + \beta \cdot x - p_i > \gamma \cdot s + \beta \cdot x - p_e$, unless both net utilities are negative, in which case consumer s does not buy at all. Given the assumptions of unit demand and uniform distribution of consumers within the interval $[0,1]$, firms' demand curves are linear and can be written as:

$$q_i = 1 - s_{ind} = 1 - \frac{p_i - p_e}{1-\gamma}; \quad q_e = s_{ind} - s_{mar} = \frac{p_i - p_e}{1-\gamma} - \frac{p_e - \beta \cdot x}{\gamma},$$

provided that the feasibility constraints $1 \geq s_{ind} \geq s_{mar} \geq 0$ hold.

Under vertical separation, the local access owner (firm *a*) sets the investment level *x* and provides wholesale access to both downstream firms (*i* and *e*) at a regulated access charge *w*. Thus, the profit functions of the three firms are the following:

$$\pi_a = w(q_i + q_e) - \frac{x^2}{2}; \quad \pi_i = p_i q_i - w q_i; \quad \pi_e = p_e q_e - w q_e.$$

Inserting for quantities, we obtain:

$$\pi_a = \left(1 - \frac{1}{1-\gamma}(p_i - p_e)\right)w + \left(\frac{p_i - p_e}{1-\gamma} - \frac{p_e - \beta \cdot x}{\gamma}\right)w - \frac{x^2}{2};$$

$$\pi_i = \left(1 - \frac{1}{1-\gamma}(p_i - p_e)\right)(p_i - w); \quad \pi_e = \left(\frac{p_i - p_e}{1-\gamma} - \frac{p_e - \beta \cdot x}{\gamma}\right)(p_e - w).$$

We define social welfare $W_S = \pi_a + \pi_i + \pi_e + CS$ as the sum of firms' profits and consumer surplus $CS = \int_{s_{ind}}^{1} (s + \beta \cdot x - p_i)ds + \int_{s_{mar}}^{s_{ind}} (\gamma \cdot s + \beta \cdot x - p_e)ds$ under vertical separation.

Solving the game backwards, first we find the third-stage optimal retail prices by the first order condition on firms' profits (given that the second order condition is always fulfilled):

$$\hat{p}_i = \frac{3w + 2(1-\gamma)(2+\beta x)}{4-\gamma}; \qquad \hat{p}_e = \frac{(2+\gamma)w + 2(1-\gamma)(\gamma + 2\beta x)}{4-\gamma}.$$

At the second stage, firm *a* maximizes its profit with respect to the investment level x. Depending on the investment spillover and the wholesale access charge (set by the regulator at the first stage), we can obtain two critical values of the investment level such that the feasibility constraints on both the marginal and the indifferent consumer are fulfilled. In such a case, we have a downstream duopoly with partial participation. However, we also find that if there is a considerable investment spillover and the access charge is sufficiently high, then the investment level is high enough to induce even the consumer located at $s = 0$ to purchase

[21] Thus, under vertical separation, the investment spillover on both downstream firms is higher than the rival firm's one, but lower than the one of firm *i*'s retail subsidiary under vertical integration.

the service from firm *e*. Conversely, when there is a limited spillover some consumers do not buy at all, and if the access charge is sufficiently high then firm *e* may stay out of the market. These results are shown in the following proposition.

Proposition 3. At the second stage of the game, depending on both the spillover effect and the wholesale access charge, the following outcomes are possible in terms of quality investment and industry structure:

i. *if both* $\beta' = \sqrt{\frac{4\gamma(1-\gamma)}{2+\gamma}} < \beta < 1$ *and* $w > w' = \frac{(4-\gamma)(1-\gamma)\gamma^2}{(2+\gamma)\left((2+\gamma)\beta^2 - (4-\gamma)\gamma\right)}$ *hold, then the marginal consumer is located at* $s = 0$ *and purchases from firm e, while the optimal quality investment is* $x_{mar}(w) = \frac{\gamma - \gamma^2 + w(2+\gamma)}{(2+\gamma)\beta}$;

ii. *if both* $\frac{2}{3} < \beta < \beta'$ *and* $w > w'' = \frac{(4-\gamma)\gamma^2}{2\left((4-\gamma)\gamma - (2+\gamma)\beta^2\right)}$ *hold, then the lowest-wtp consumers do not buy, firm i has a downstream monopoly, and the optimal quality investment is* $x_{ind}(w) = \frac{(2w-\gamma)}{2\beta}$;

iii. *either if both* $\beta' < \beta < 1$ *and* $0 < w < w'$ *hold, or if both* $\frac{2}{3} < \beta < \beta'$ *and* $0 < w \le w''$ *hold, then the lowest-wtp consumers do not buy, both downstream firms are active, and the optimal quality investment is* $x_{int}(w) = \frac{w\beta(2+\gamma)}{(4-\gamma)\gamma}$.

Proof. See Appendix.

Solving the regulator's problem at the first stage of the game we find that, contrary to the integrated scenario, vertical separation rules out the possibility that one of the retail firms stays out of the market. Hence, there is always a downstream duopoly. In the case when the investment spillover is limited (or, the efficiency loss due to vertical disintegration is large), we find an interior solution where both the overall quantity sold and quality investment rise with the spillover effect, that is, $\frac{\partial(q_i + q_e)}{\partial\beta} > 0$, and $\frac{\partial x_{int}^a}{\partial\beta} > 0$. On the other hand, when the investment spillover is large (or, the efficiency loss due to vertical disintegration is limited), we find a frontier solution where the lowest-wtp consumer purchases from firm *e*. Proposition

4 below formalizes the results, while the following Table 2 summarizes the outcome of the game in terms of firms' market shares, quality investment, access charge, and social welfare.

Proposition 4. Under vertical separation, there is always a downstream duopoly. If $\beta'' < \beta < 1$, *then the regulator sets* $w = w_{mar}^a$, *and the consumer located at* $s = 0$ *buys from firm e. If* $\frac{2}{3} < \beta \leq \beta''$ *then the regulator sets* $w = w_{int}^a$ *and the lowest-wtp consumers do not buy at all.*

Proof. See the Appendix, where we relegate the expression of β''.

	$\frac{2}{3} < \beta \leq \beta''$	$\beta'' < \beta < 1$
q_i	$\frac{3\beta^4(-2+\gamma)(2+\gamma)^2+2(4-\gamma)^2(1+2\gamma)\gamma^2+\gamma(4-\gamma)(2+\gamma)(8-5\gamma)\beta^2}{(-4+\gamma)^2\gamma^2 4(2+5\gamma)-(2+\gamma)^2(12-\gamma-2\gamma^2)\beta^4-\gamma(2+\gamma)(4-\gamma)\beta^2(16+\gamma(4+3\gamma))}$	$\frac{1+\gamma}{2+\gamma}$
q_e	$\frac{3(-4+\gamma)^2\gamma^2-3\gamma(2-\gamma)(4-\gamma)(2+\gamma)\beta^2+(2+\gamma)^2(5-2\gamma)\beta^4}{(-4+\gamma)^2\gamma^2 4(2+5\gamma)-(2+\gamma)^2(12-\gamma-2\gamma^2)\beta^4-\gamma(2+\gamma)(4-\gamma)\beta^2(16+\gamma(4+3\gamma))}$	$\frac{1}{2+\gamma}$
x	$x_{int}^a = \frac{\gamma(4(4-\gamma)(1-\gamma)\gamma-(2+\gamma)(16-7\gamma)\beta^2)\beta}{(-4+\gamma)^2\gamma^2 4(2+5\gamma)-(2+\gamma)^2(12-\gamma-2\gamma^2)\beta^4-\gamma(2+\gamma)(4-\gamma)\beta^2(16+\gamma(4+3\gamma))}$	$x_{mar}^a = \beta$
w	$w_{int}^a = \frac{(-4+\gamma)\gamma^2(4(4-\gamma)(1-\gamma)\gamma-(2+\gamma)(16-7\gamma)\beta^2)}{(-4+\gamma)^2\gamma^2 4(2+5\gamma)-(2+\gamma)^2(12-\gamma-2\gamma^2)\beta^4-\gamma(2+\gamma)(4-\gamma)\beta^2(16+\gamma(4+3\gamma))}$	$w_{mar}^a = \beta^2+\gamma-3+\frac{6}{2+\gamma}$
W_s	$\frac{3(-4+\gamma)^2\gamma^2(1+2\gamma)+(2+\gamma)^2(1-\gamma)(4\gamma-9)\beta^4+\gamma\beta^2(32+\gamma(-8+\gamma(2+(3-2\gamma)\gamma)))}{2((-4+\gamma)^2\gamma^2 4(2+5\gamma)-(2+\gamma)^2(12-\gamma-2\gamma^2)\beta^4-\gamma(2+\gamma)(4-\gamma)\beta^2(16+\gamma(4+3\gamma)))}$	$\frac{3+\beta^2(2+\gamma)^2(5+\gamma)\gamma}{2(2+\gamma)^2}$

Table 2. Vertical separation – outcome of the game.

3.3 Welfare analysis

Let us now compare the outcomes of the game in the two alternative scenarios of vertical integration and separation of the local access owner, both in terms of quality investment and social welfare. In what follows, we analyze and discuss the main results.

First, we find the expected result that vertical separation generally induces higher downstream competition than vertical integration. While under vertical separation the downstream market is always a duopoly, in some circumstances the vertically integrated firm invests so much as to achieve a downstream monopoly (even though the wholesale access charge is regulated). More interestingly, we find that vertical separation does not necessarily face the trade-off between promoting downstream competition and ensuring network quality investment. Indeed, our results show that it is perfectly possible that whenever separation

induces higher downstream competition than integration, it also induces higher investment in network quality. The results obtained also show that both quality investment and social welfare benefit from vertical separation in any of the following circumstances: a) the investment spillover is sufficiently high; b) the (perceived) quality of the rival firm's product is sufficiently high; c) the investment spillover and the rival firm's (perceived) quality are sufficiently asymmetric.

The following proposition proves the results in the case when under vertical separation downstream firms have a considerable ability to use quality investment (so that separation induces even the lowest-wtp consumer to purchase from firm *e*).

Proposition 5. Let $\beta > \beta''$. *If* $\delta'(\gamma) \leq \delta < 1$ *then: i)* $x^a_{mar} > x^i_{mar}$ *when* $\beta > x^i_{mar} > \beta''$, *and ii)* $W_s > W_I$ *when* $\beta > \hat{\beta} > \beta''$; *else, if* $\frac{2}{3} \leq \delta < \delta'(\gamma)$ *then: iii)* $x^a_{mar} > x^i_{ind}$ *when* $\beta > x^i_{ind} > \beta''$, *and iv)* $W_s > W_I$ *if* $\beta > \tilde{\beta} > \beta''$.

Proof. See Appendix, where we relegate the expressions of $\hat{\beta}$ and $\tilde{\beta}$.

The same qualitative results of Proposition 5 do hold when the downstream firms' ability to use quality investment under separation is not so high (i.e. when $\beta \leq \beta''$), so that the lowest-wtp consumers are not active. Thus, both quality investment and welfare improve with vertical separation if the spillover effect is not too low. Although we do not provide the analytical results here (since the expressions of both investment and welfare are too complicated), we carry out a numerical analysis to illustrate the point at issue, where the downstream firms' ability to exploit the upstream quality investment is set at three different (increasing) values ($\beta_l = \delta$, $\beta_m = \frac{1+\delta}{2}$, and $\beta_h = \frac{4+\delta}{5}$, where subscripts *l*, *m*, and *h* respectively stand for low, medium and high ability). Note that the case when $\beta_l = \delta$ is the worst case for vertical separation, given that both downstream firms' ability is equal to the investment spillover to firm *e* under vertical integration, and thus is the case associated with the highest economies of integration.

Figure 1 compares quality investment in the two alternative scenarios, while Figure 2 compares social welfare. Depending on the rival firm's (perceived) service quality γ and the investment spillover δ (or β), three different situations may arise: (i) a downstream duopoly where the lowest-wtp consumer buys from firm *e*, both under integration and separation (over

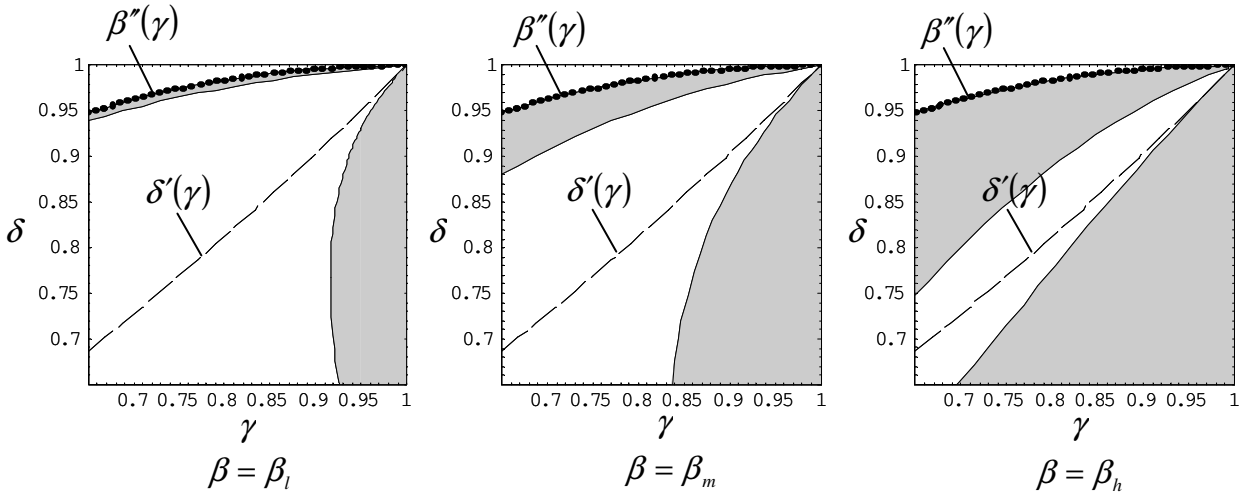


Figure 1. Separation vs. integration - A comparison of quality investment.

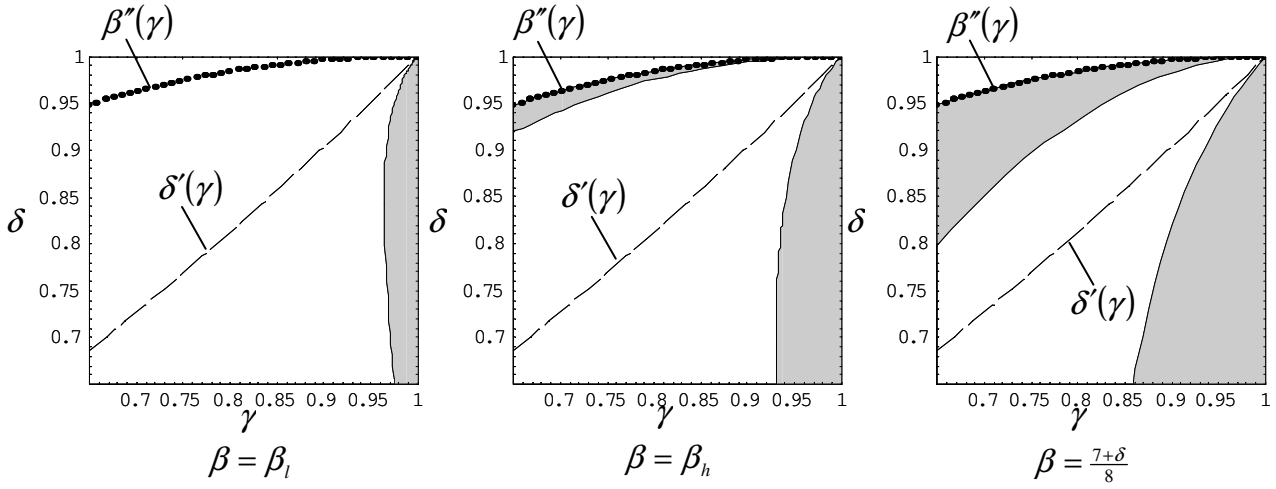


Figure 2. Separation vs. integration - A comparison of social welfare.

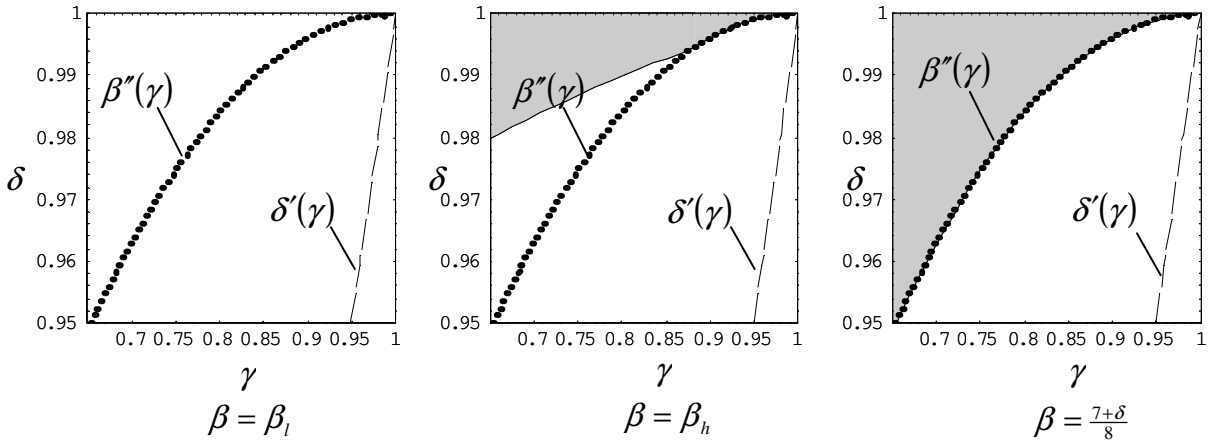


Figure 3. Separation vs. integration - A comparison of quality investment.

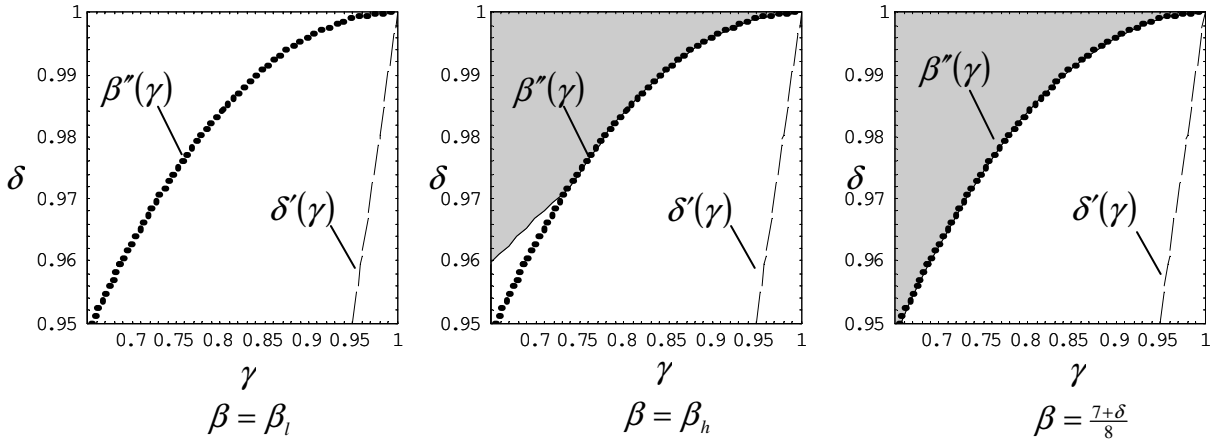


Figure 4. Separation vs. integration - A comparison of social welfare.

the dotted line), (ii) a downstream duopoly where the marginal consumer is located at $s = 0$ in the integrated scenario, whereas the lowest-wtp consumers are not active under separation (between the dashed line and the dotted line), (iii) a vertically integrated monopoly where the lowest-wtp consumer buys from firm *i* versus a downstream duopoly where the lowest-wtp consumers are not active under separation (below the dashed line). We find that within the shaded areas both quality investment (Figure 1) and social welfare (Figure 2) under separation is higher than under integration. Thus, the results shown in Figure 1 and Figure 2 confirm what stated at points a), b), and c) above.

Figure 3 and 4 respectively compare investment and welfare in both scenarios within the exclusive portion of the (δ, γ) plan where the results of Proposition 5 hold[22], and consider three different cases (where β is respectively equal to β_m, β_h, and $\frac{7+\delta}{8} > \beta_h$).

The results obtained by the numerical analysis and those deriving from Proposition 5 point out a unique trend that relates both quality investment and welfare to the spillover effect. In fact, the larger the difference between the spillover effect under separation and integration (i.e. the larger $\beta - \delta$), the higher both quality investment and welfare under separation than integration. This means that separation is particularly effective when the vertically integrated firm significantly reduces the input quality to the rival firm. What is essential to remark here is that ensuring equivalent access to downstream competitors through vertical separation also produces higher investment in the quality of the access network.

4 Concluding remarks

Mandatory vertical separation of the dominant firm in fixed telecommunications can be both an effective and proportionate regulatory option to prevent price and non-price discrimination of downstream competitors, particularly in those countries where (i) the local access network is an enduring economic bottleneck, so that both within-platform and between-platform end-to-end competition in the mass market is not sustainable, and (ii) the vertically integrated firm has repeatedly breached either the regulatory contract or antitrust laws.

One of the most critical issues is the common presumption that vertical (either functional or structural) separation of the local access owner would cause a decline in network quality

[22] This portion is not explicitly analyzed in Figure 1 and Figure 2.

investments, which has been supported by some literature findings relevant to specific network industries (such as energy and railways). Given that several incumbents worldwide have recently announced or undertaken the deployment of NGANs, it is meaningful to assess whether or not such a presumption does hold as far as fixed telecommunications is concerned.

In this framework, we have shown that this presumption can be incorrect when we consider two basic features of the retail broadband access market, that is, imperfect price competition with differentiated products and partial market participation. It follows that quality-enhancing network investment is higher under vertical separation than integration, provided that downstream firms have a sufficiently high ability to offer advanced services on the basis of such investment, or the perceived quality differential between the incumbent's retail product and the rival firm's one (e.g. due to consumer switching costs) is sufficiently low. We have found that mandatory separation improves network investment when the vertically integrated firm is far from ensuring equivalent access to downstream competitors. We have also found that an increase in quality investment related to vertical separation mostly occurs in conjunction with an improvement in social welfare.

We have obtained these results albeit our model includes efficiency losses due to vertical disintegration. In fact, the incumbent's disintegration is usually associated with diseconomies of scale and scope, as well as coordination problems among upstream and downstream firms. It is however worth noting that, even with vertical integration diseconomies of scale and scope are "imposed" by market liberalization. If social gains from competition are deemed to outweigh costs, then in several circumstances mandatory vertical separation is the most effective means to promote competition, which in turn may encourage network investment.

The results obtained depend in part on model formulation, and thus on the specific assumptions on demand and cost functions, as well as the nature of downstream competition. Nonetheless, the qualitative result that vertical separation may raise quality investment is not diluted, but rather strengthened within a number of alternative model specifications.

First, the direct effect of vertical separation is the creation of a level playing field for downstream competition. We model this effect by removing the assumption that the incumbent has a superior ability to offer value-added services on the basis of network investment, which does hold under vertical integration. Clearly, this is well suited to emphasize that input quality discrimination can no longer be applied under separation. However, we rule out the case when downstream competition induced by separation is such that (i) investment in network quality specifically enables the rival firm to offer differentiated services and applications from the incumbent, (ii) the rival firm's services induce a higher

increase in consumer wtp than the incumbent, (iii) both firms simultaneously benefit from network investment in such a way that the provision of advanced services does offset efficiency losses due to vertical disintegration. It follows from the foregoing statements that we are undervaluing the positive effect that separation may have on network investment.

Second, the proposed model is one of vertical differentiation, and as such prescribes that even under separation consumers with high wtp for broadband services buy from the incumbent, while the rival firm has to reduce the retail price to attract new low-spending users. Although this may indeed hold shortly after the incumbent's disintegration, it may not necessarily hold when vertical separation has fully deployed its pro-competitive effect. In such a case, a model of horizontal differentiation *à la* Hotelling could better fit downstream competition under separation. Since horizontal differentiation lightens retail price competition, then it also has a positive impact on upstream investment.

Third, we have set the degree of convexity of the investment cost function at such a level that in a number of circumstances the vertically integrated firm optimally raises quality investment to foreclose the downstream market (without any prejudice to market participation), given that the rival firm has a limited investment spillover. If the investment cost function is sufficiently convex, then *ceteris paribus* the integrated firm would not always find it profitable to increase quality investment to the extent that is necessary to achieve a downstream monopoly. In this sense, if investment is more costly then it is plausible that network quality decreases more under vertical integration than separation.

Finally, there are several dimensions along which our research can be extended in future work. First, we can refine our analysis to assess how different forms of vertical separation (ranging from simple accounting separation, through functional separation to full structural or ownership separation) affect the potential trade-off between promoting competition and ensuring the incumbent's investment. Second, there is the risk that the deployment of NGANs amplifies the digital divide between the most and the least developed areas of the country. If wide broadband adoption is a primary policy goal, then there is the question whether or not broadband access should be part of universal service obligations. This in turn raises the need to fine-tune the current universal service funding model, and poses the question whether and how public intervention may bridge the broadband gap. In this framework, public-funded projects targeting passive infrastructure (such as ducts, dark fibre, or antenna sites) may be well suited. However, it is essential to ensure open access to such infrastructure to lower entry barriers and increase investment incentives, while not distorting competition. In this sense, vertical separation of the local access owner may still be an appropriate remedy.

Appendix

Proof of Proposition 1. Firm *i*'s profit function is strictly concave with respect to quality investment *x*. Hence, by the first order condition on firm *i*'s profit we find the optimal investment level $x_{int}(w)=\frac{(1-\gamma)\left((4-w)\gamma^2-4\gamma(2-\delta)+8\delta w\right)}{\gamma\left(-7\gamma^2+\gamma^3+4\gamma(4+\delta)-2(4+(4-\delta)\delta)\right)}$. However, $x_{int}(w)$ is a feasible solution if and only if $1\geq s_{ind}(x_{int}(w))\geq s_{mar}(x_{int}(w))\geq 0$. Computation yields that this chain of conditions hold if and only if both $\gamma\leq\delta\leq\frac{1}{2}(3\gamma-\gamma^2)$ and $w'\leq w\leq w''$ hold, where $w'=\frac{\gamma\left((-3+\gamma)\gamma^2-2(\delta-2)\delta\right)}{2\gamma^3-4\delta^2(7+\delta)+\gamma(4+6\delta)}$ and $w''=\frac{\gamma\left((-3+\gamma)(\gamma-2)\gamma+4(\gamma-2)\delta+2\delta^2\right)}{-4\gamma^2+\gamma^3+2\delta(\delta-2)+\gamma(2+3\delta)}$.

In the remaining cases, either $s_{mar}(x_{int}(w))\geq 0$ or $s_{ind}(x_{int}(w))\geq 0$ does not hold. First, let $s_{mar}(x_{int}(w))<0$ but $s_{ind}(x_{int}(w))\geq 0$. In such a case, we find the optimal investment by solving the equation $s_{mar}(x(w))=0$, and obtain $x_{mar}(w)=\frac{\gamma-\gamma^2+w(2+\gamma)^2}{\gamma+2\delta}$. It is easy to verify that $s_{ind}(x_{mar}(w))\geq 0$ if and only if $w\leq w_{ind}=\frac{(1-\gamma)\delta}{1-\delta}$. If $w>w_{ind}$, then we have that $s_{ind}(x_{mar}(w))<0$. In such a case, we find the optimal investment by solving the equation $s_{ind}(x(w))=0$, and obtain $x_i=x_{ind}(w)=\frac{(1-\gamma)(2+w-\gamma)}{2-\gamma-\delta}$. ■

Proof of Proposition 2. Proposition 1 shows that, if both $\gamma\leq\delta\leq\frac{1}{2}(3\gamma-\gamma^2)$ and $w'\leq w\leq w''$ hold then the optimal quality investment is $x_{int}(w)$. Let $x=x_{int}(w)$. Since the welfare function is concave in w then the regulator maximizes welfare by the first order condition $\left.\frac{\partial W(w)}{\partial w}\right|_{x=x_{int}(w)}=0$, and finds the access charge w_{int} (for brevity, we omit the expression of w_{int}). However, computation yields that w_{int} is such that $w'\leq w_{int}\leq w''$ cannot hold, so that the feasibility constraints $1\geq s_{ind}(x_{int}(w_{int}))\geq s_{mar}(x_{int}(w_{int}))\geq 0$ cannot be fulfilled. In such a case, the consumer in $s=0$ buys from firm *e* and the optimal investment level is $x_{mar}(w)$ if and only if the optimal access charge is lower than w_{ind}. Inserting for $x_{mar}(w)$ and solving for

$\left.\frac{\partial W(w)}{\partial w}\right|_{x=x_{mar}(w)}=0$, we find the access charge $w_{mar}=\frac{(\gamma-1)\left(\gamma(1+\gamma)^2-(-3+\gamma)\delta+(1+2\gamma)\delta^2\right)}{-1+3\gamma^2+\gamma^3+2\gamma(-1+\delta)^2+3(-2+\delta)\delta}$.

With some algebra, we obtain that $w_{mar}\le w_{ind}$ holds if $\frac{4+\gamma}{2}-\frac{\sqrt{3(4-\gamma^2)}}{2}=\delta'(\gamma)\le\delta<1$.

In the case when $\frac{2}{3}<\delta<\delta'(\gamma)$, we have that $s_{ind}(x_{mar}(w_{mar}))<0$. Hence, the consumer in $s=0$ buys from firm *i* and the optimal investment is $x_i=x_{ind}(w)$. Since $s_{ind}(x_{ind}(w))=0$, then the condition $s_{ind}(x_{ind}(w))\ge s_{mar}(x_{ind}(w))\ge 0$ is binding. It follows that the optimal access charge w_{ind} is obtained by solving the equation $s_{ind}(x_{ind}(w))=s_{mar}(x_{ind}(w))$. ■

Proof of Proposition 3. Since $\pi_a(x,w)$ is strictly concave with respect to x, then by the first order condition we obtain the optimal quality investment $x_{int}(w)=\frac{w\beta(2+\gamma)}{(4-\gamma)\gamma}$. Computation yields that, if both $\beta'=\sqrt{\frac{4\gamma(1-\gamma)}{2+\gamma}}<\beta<1$ and $w>w'=\frac{(4-\gamma)(1-\gamma)\gamma^2}{(2+\gamma)\left((2+\gamma)\beta^2-(4-\gamma)\gamma\right)}$ hold, then $s_{mar}(x_{int}(w))<0$. It follows that the constraint on the marginal consumer is binding, and the optimal quality investment $x_{mar}(w)=\frac{\gamma-\gamma^2+w(2+\gamma)}{(2+\gamma)\beta}$ derives from solving the equation $s_{mar}(x(w))=0$. Alternatively, if both $\frac{2}{3}<\beta<\beta'$ and $w>w''=\frac{(4-\gamma)\gamma^2}{2\left((4-\gamma)\gamma-(2+\gamma)\beta^2\right)}$ hold, then we find that $s_{ind}(x_{int}(w))\ge s_{mar}(x_{int}(w))$ is not fulfilled. Given that $s_{mar}(x_{int}(w))>0$, the optimal investment $x_{ind}(w)=\frac{(2w-\gamma)}{2\beta}$ derives from solving $s_{ind}(x(w))=s_{mar}(x(w))$. In all of the remaining cases, since $1\ge s_{ind}(x_{int}(w))\ge s_{mar}(x_{int}(w))\ge 0$ holds then $x_{int}(w)$ is optimal. ■

Proof of Proposition 4. Assume that the feasibility constraints on both the marginal and the indifferent consumer are not violated. Thus, case *iii.* of Proposition 3 states that the optimal investment is $x_{int}(w)$. Given that social welfare is a strictly concave function in *w*, solving $\left.\frac{\partial W(w)}{\partial w}\right|_{x=x_{int}(w)}=0$ we obtain $w_{int}=\frac{(-4+\gamma)\gamma^2\left(4(4-\gamma)(1-\gamma)\gamma-(2+\gamma)(16-7\gamma)\beta^2\right)}{(-4+\gamma)^2\gamma^2 4(2+5\gamma)-(2+\gamma)^2\left(12-\gamma-2\gamma^2\right)\beta^4-\gamma(2+\gamma)(4-\gamma)\beta^2(16+\gamma(4+3\gamma))}$.

Computation yields that $w_{int}\le w'$ if and only if $\beta\le\beta''=\sqrt{\frac{(4-\gamma)\gamma\left(14-2\gamma-3\gamma^2+\sqrt{2+\gamma}\sqrt{76-44\gamma-14\gamma^2+9\gamma^3}\right)}{44+22\gamma-8\gamma^2-4\gamma^3}}$,

where $\beta'' > \beta'$ always holds. Hence, if $\beta' < \beta \leq \beta''$ then w_{int} is the optimal access charge. If $\frac{2}{3} < \beta < \beta'$, then we find that $w_{int} < w''$ always holds, so that w_{int} is still optimal, whereas case *ii.* in Proposition 3 is ruled out. Conversely, if $\beta > \beta''$ then $w_{int} > w'$ and the optimal investment is $x_{mar}(w)$, as derived in case *i.* of Proposition 3. Hence, the optimal access charge is found by solving $\left.\frac{\partial W(w)}{\partial w}\right|_{x=x_{mar}(w)} = 0$, and is $w_{mar} = \beta^2 + \gamma - 3 + \frac{6}{2+\gamma}$. ■

Proof of Proposition 5. As regards quality investment, the proof directly follows from the expressions of the optimal investments respectively shown in Table 1 and in Table 2. As regards social welfare, computation yields that welfare improves under separation (i.e. $W_s > W_I$) in the following cases: i) when $\delta'(\gamma) \leq \delta < 1$, provided that $1 > \beta > \hat{\beta} = \sqrt{\frac{(\gamma-1)(3+\delta+\gamma(3+\delta+\gamma))^2}{(2+\gamma)^2(-1+3\gamma^2+\gamma^3+2\gamma(-1+\delta)^2+3(-2+\delta)\delta)}}$; ii) when $\frac{2}{3} \leq \delta < \delta'(\gamma)$, provided that $1 > \beta > \tilde{\beta} = \sqrt{\frac{(1-\gamma)(5+8\gamma+5\gamma^2+\gamma^3-2(5+\gamma(4+\gamma))\delta+\delta^2)}{(2+\gamma)^2(1-\delta)^2}}$. ■

References

Analysys (2007), "The business case for sub-loop unbundling in the Netherlands", Final Report for OPTA, 26 January 2007.

Armstrong M. (2002), "The theory of access pricing and interconnection", in: Cave M., S. Majumdar and I. Vogelsang (eds.), *Handbook of Telecommunications Economics*, vol. 1, North Holland, Amsterdam, 295-384.

Bourreau M. and P. Dogan (2005), "Unbundling the local loop", *European Economic Review* 49, 173-199.

Bourreau M. and P. Dogan (2006), "Build-or-buy strategies in the local loop", *American Economic Review* 96, 72-76.

Buehler S., A. Schmutzler and M. Benz (2004), "Infrastructure quality in deregulated industries: is there an underinvestment problem?", *International Journal of Industrial Organization* 22, 253-267.

Buehler S., D. Gartner and D. Halbheer (2006), "Deregulating network industries: dealing with price-quality tradeoffs", *Journal of Regulatory Economics* 30, 99-115.

Cave M. (2006), "Six degrees of separation. Operational separation as a remedy in European telecommunications regulation", *Communications & Strategies* 64, 1-15.

Cave M. and I. Vogelsang (2003), "How access pricing and entry interact", *Telecommunications Policy* 27, 717-727.

COCOM – Communications Committee (2007), *Broadband access in the EU: situation at 1 January 2007*, European Commission, COCOM 07-36 FINAL, July 2007.

de Bijl P. (2004a), "Structural separation and access in telecommunications markets", TILEC Discussion Paper DP 2004-011, Tilburg University.

de Bijl P. and M. Peitz (2004b), "Dynamic regulation and entry in telecommunications markets: a policy framework", *Information Economics and Policy* 16, 411-437.

ERG – European Regulators Group (2006), *Revised ERG Common Position on the approach to appropriate remedies in the ECNS regulatory framework (Final Version)*, available at: http://erg.eu.int.

ERG – European Regulators Group (2007), *ERG Opinion on functional separation*, ERG (07) 44, available at: http://erg.eu.int.

Faulhaber G.R. (2003), "Policy-induced competition: the telecommunications experiments", *Information Economics and Policy* 15, 73-97.

Foros O. (2004), "Strategic investments with spillovers, vertical integration and foreclosure in the broadband access market", *International Journal of Industrial Organization* 22, 1-24.

Gans J.S. and S.P. King (2004), "Access holidays and the timing of infrastructure investment", *Economic Record* 80, 89-100.

Mandy D.M. and D.E.M. Sappington (2007), "Incentives for sabotage in vertically related industries", *Journal of Regulatory Economics* 31, 235-260.

OECD (2003), "The benefits and costs of structural separation", TISP Paper on Competition and Regulation, DAFFE/COMP/WP2(2003)2.

OECD (2006), *Report to the Council on experiences on the implementation of the recommendation concerning structural separation in regulated industries*, C(2006)65.

OFCOM (2005), *Final statements on the Strategic Review of Telecommunications, and undertakings in lieu of a reference under the Enterprise Act 2002*, available at: http://www.ofcom.org.uk.

Peitz M. (2005), "Asymmetric access price regulation in telecommunications markets", *European Economic Review* 49, 341-358.

Pindyck R.S. (2007), "Mandatory unbundling and irreversible investment in telecom networks", *Review of Network Economics* 6, 274-298.

Valletti T. and C. Cambini (2005), "Investments and network competition", *RAND Journal of Economics* 36, 446-467.

Information Technology, efficiency and productivity in SMEs: Evidence for Portugal

María Rosalía Vicente Cuervo* and Maria do Rosário O. Martins**

Abstract

This study examines how information technology (IT) contributes to the performance of small and medium-sized enterprises (SMEs) using a sample of Portuguese firms and considering two measures of firm performance, technical efficiency and productivity. The study is different from received analysis in that it takes into account the effects of both IT capital and the intensity of IT use. The empirical findings show that IT capital itself does not enhance Portuguese effectiveness. Meanwhile IT adoption and use has a positive impact on performance, especially when complemented with a high-qualified labour force, as well appropriate management practices.

Key words: information technology; small and medium-sized enterprises (SMEs); technical efficiency; productivity

JEL codes: D2, L2

* Department of Applied Economics, Campus del Cristo s/n, University of Oviedo, 33006-Asturias, Spain. Tel. + 34 985 10 50 53; fax. + 34 985 10 50 50. E-mail address: mrosalia@uniovi.es.

** Instituto Superior Estatística e Gestão de Informação, Universidade Nova de Lisboa. Campus de Campolide 1070-312 Lisboa, Portugal. E-mail address: mrfom@isegi.unl.pt

1. Introduction

During the last decade the management of information technology (IT) has become one of the critical issues for business managers to address. The impact of IT has been perceived in almost every part of a business, creating advantages from better access to knowledge and information, lower transaction costs, coverage of larger markets, improved decision-making, greater flexibility in catering to a diversified customer base, and increased overall productivity (Song and Mueller-Falcke, 2006).

Within this context the analysis of the economic impact of IT has been a subject of intense investigation over the last years. Dedrick et al. (2003) review more than fifty papers that explore the impact of IT on productivity at country, industry and firm level. More recently, Dracca et al. (2007) include a survey of almost forty articles.

Three comments on this growing body of literature deserve special attention. First, most studies examine the effects of IT on the change in labour or total factor productivity, and do not typically consider other performance measures, such as technical efficiency. Second, IT-related equipment expenditure is usually employed to proxy for the state of IT. While IT stock may provide an accurate measure of IT investment, it is important to take into account that firm performance is also affected by the intensity and application of IT use (Sung, 2007). And third, there are relatively few studies that address the analysis of IT effects on the performance of small and medium-sized enterprises (SMEs) (Morikawa, 2004; Shin, 2006; Johnston et al., 2007). This last point is especially unfortunate since these technologies hold a large potential for SMEs, such as the provision of the opportunity to overcome the limitations of size and to compete effectively in larger markets with bigger sized establishments (Dholakia and Kshetri, 2004; Lucchetti and Sterlacchini, 2004). However, IT access and investments

alone are no panacea for SMEs. Unless they are accompanied by organizational changes and innovative ideas, SMEs might end up under-utilizing the available technologies.

The aim of this paper is to overcome the outlined gaps in the specialised literature by examining the impact of IT on the technical efficiency of a sample of Portuguese SMEs.

It is important to note that we attempt to take into account the effects of both IT capital and the intensity of IT use, compared to previous studies which have focused on the former. The efficiency analysis is carried out in two stages. The first stage involves the use of data envelopment analysis (DEA) to measure the levels of technical efficiency. In the second stage, efficiency levels are treated as a dependent variable and regressed upon the corresponding IT variables to examine whether they have a positive influence on technical efficiency. Since technical efficiency and productivity are two concepts closely related, we also analyse the effect of IT on productivity for comparison purposes.

SMEs play a significant role in Portuguese economy since they comprise 99% of the enterprises, account for 68.1% of the added value and 82.2% of the total employment. (INE, 2005). Nonetheless, the productivity of Portuguese SMEs is one of the lowest in the European Union (OECD, 2006).

With regard to IT use among Portuguese SMEs, official statistics show that computer penetration rate is high (95%). However, computer usage by workers is still relatively low: only 30% of workers in Portuguese SMEs regularly use computers in their tasks (INE, 2005).

In these domestic circumstances, analyzing whether IT has real effects on business performance and what effects, if any, becomes a crucial issue. To out

knowledge this is the first attempt to show evidence for Portuguese SMEs and results may be helpful in order to understand IT's impacts on SMEs in other countries with similar business structure such as those in Southern Europe (Spain, Italy and Greece).

The paper is organised as follows. Section 2 presents the background on the relationship between IT, efficiency and productivity. Section 3 describes the methodology used for the efficiency analysis. Section 4 offers a description of the database examined. Section 5 presents the empirical results and their interpretation, and we finally draw some concluding remarks.

2. IT, efficiency, and productivity

Technical efficiency is an important and useful economic measure of organizational performance, which is closely related to, but different from, productivity (Lin and Shao, 2000, 2006).

On one hand, productivity relates the amount of outputs produced to the amount of inputs consumed and is defined as the ratio of aggregate outputs to aggregate inputs.

On the other hand, technical efficiency reflects the ability of a firm either to produce the maximum output attainable from a given set of inputs or to use possible minimum input to produce a given set of outputs.

Using the maximum output criterion, assume that a firm operates at a point of the production possibility set that represents the set of all technologically feasible production plans for a given level of inputs. The corresponding production frontier is the maximum output attainable from a given set of inputs. A firm is technically efficient only when it operates on the production frontier. Therefore, the technical efficiency of a

firm is measured by comparing its output level with that of the technically efficient or best firms for the given input level.

A crucial connection between technical efficiency and productivity can be established: productivity growth is a composite index of the change in technical efficiency and the shift in the production frontiers (technical change).

In this framework, technical inefficiency may result from a number of causes that would unfavourably influence a firm's ability to use its input resources fully in producing output. Some of these undesirable causes are beyond the firm's control (weather, accidents, and strikes). Meanwhile, others are attributable to the firm itself and can be corrected through its efforts to improve the production process; such are the cases of ineffective communication and poor decision-making, among others (Shao and Lin, 2002).

The integration of IT in the production process may eliminate or, at least, reduce some of these inefficiencies.

Thus, seamless transfer of information through shared electronic files and networked computers increases the efficiency of business processes such as documentation, data processing and other back-office functions. At the same time, IT networks help to transfer messages quickly among the employees in an organization, overcoming distance and time zone differences, and facilitating effective communication.

Moreover, increasingly sophisticated IT applications such as CRM (Customer Relation Management), ERP (Enterprise Resource Planning), and KMS (Knowledge Management System) allow firms to store, share and use their acquired knowledge and know-how, helping managers in the process of decision-making (OECD, 2004). In

particular, CRM not only provides the latest client-related information, which better equips managers and employees for responding to customer inquiries, but also makes business processes and knowledge accumulation more efficient. Additionally, IT may help reduce inefficiency in the use of capital and labour by reducing inventories, and the more customers or firms are connected to the network, the greater the benefits (spillover effects) (OECD, 2002).

Therefore, IT is expected to enhance an organization's ability to produce more output using the same amount of inputs, or alternatively produce the same output using less input. In other words, IT is expected to improve a firm's technical efficiency, and through it, its productivity (given the aforementioned connection between these two measures).

Nonetheless, it is important to take into account that to accomplish such improvements, the use and application of IT require other complementary factors. In this sense, previous evidence (Arvanitis, 2005; Black and Lynch, 2001; Bresnahan et al., 2002; Brynjolfsson and Hitt, 1996, 2003; OECD, 2003; Zwick, 2003) shows that a well-educated labour force is a basic element to use IT efficiently and make computers and related technologies more productive. Furthermore, investments in skills, organisational change and innovation are key elements to make IT work. Without these, the economic impact of IT at firm level may be limited.

3. Research methods

In this paper, the empirical analysis of technical efficiency is divided in two stages. In the first stage, technical efficiency of each individual firm is measured by means of data envelopment analysis (DEA), a non-parametric technique that is well

known in the field of operations research and that has been successfully employed to explore the role of IT in technical efficiency and productivity as shown by Milana and Zeli (2004) in a paper for the OECD. In the second stage, the correlation between IT and technical efficiency is examined using cross-sectional Tobit regressions run on firm-level data.

3.1. Data envelopment analysis: measuring firms' technical efficiency

A systematic measure of technical efficiency was first constructed by Farrell (1957), and two major approaches for evaluating efficiency were subsequently developed: parametric and non parametric methods.

The parametric approach requires the assumption of a functional form (Cobb–Douglas, translog, CES, etc.) to be made for the production frontier; it uses the statistical estimation to estimate the coefficients of the production function as well as the technical efficiency (Lowell, 1993). Since the parametric production frontier is assumed to be the ''true'' frontier, the scores of technical efficiency obtained are regarded as absolute technical efficiency.

Nonparametric production frontiers, on the other hand, are based on mathematical programming and do not make any assumptions about the functional form. The data points in the data set are compared with one another for efficiency. The most efficient observations are utilized to construct the piece-wise linear convex nonparametric frontier. As a result, nonparametric production frontiers are employed to measure relative technical efficiency among the observations. Perhaps, the most popular non parametric method is data envelopment analysis (DEA), launched by Charnes, Cooper and Rhodes (1978; CCR model for short) under the assumption that production exhibit constant returns to scale. Banker, Charnes and Cooper (1984, BCC model for short) extended it to the case where there are variable returns to scale.

DEA has some advantages over parametric methods. First, it is not necessary to assume some functional form for the production function, and therefore, there is no risk of misspecifying it. Second, it is relatively insensitive to model specification, the efficiency measurement is similar if it is estimated oriented to inputs or oriented to outputs[1]. Third, it enables us to identify "best practice" units, that is, those that are on the frontier, and provides information of the nature of the inefficiency of a particular unit. Fourth, it can handle multiple outputs and inputs

Consider N decision-making-units (DMUs) and let the inputs and outputs of the k^{th} DMU be respectively represented by the input vector x_k and output vector y_k. Input and output data for all DMUs forms the input matrix X and output matrix Y.

Assuming the general case, which includes variable returns-to-scale, the basic DEA model (BCC model) is expressed by the linear programming problem:

$$\begin{aligned} \text{Max}_{\theta,\lambda} \quad & \varphi \\ \text{s.t.} \quad & x_k - X\lambda \geq 0 \\ & -\varphi y_k + Y\lambda \geq 0 \\ & N1\lambda = 1 \\ & \lambda \geq 0 \end{aligned} \qquad (1)$$

With φ being a scalar, λ is an N-order column vector of constants, N1 is an N-order column vector of ones. The convexity condition, $N1\lambda=1$, ensures that an inefficient DMU is only benchmarked against DMUs of a similar size[2].

[1] Measures of efficiency rely on estimating maximum output for given levels of inputs (output orientation), or alternatively minimum inputs level for given output levels (input orientation).

[2] In the case of constant returns-to-scale (CCR model), this constraint is not imposed, so λ weights sum up to a value different from one and the benchmarking may be made against DMUs that are substantially larger or smaller than the examined DMU.

The optimal solution of this problem, φ^*, represents the proportional increase in outputs that could be achieved by the k^{th} DMU, with input levels being constant. In this sense, this DEA model is output-oriented and φ^* is equivalent to the technical efficiency score. If this score is equal to one, the DMU is on the frontier and therefore, efficient; if the score is greater than one, then the DMU is said to be inefficient. In this paper, the DEA results are reported as $1/\varphi$ for an easy interpretation as a percentage.

The BCC model (assuming variable returns-to-scale) allows efficiency scores to be separated into two components: on one hand, "pure technical efficiency" (or "managerial efficiency") which is input based and measures whether the firm is using too much input to produce a given level of output; and, on the other hand, "scale efficiency" which is output based and it determines whether the firm is operating on the right scale or not.

It is worth noting some caveats associated with DEA (Charnes et al., 1994). Since it infers the best practices production function from the reported input-output combinations of some small number of the most efficient firms, results may be highly sensitive to measurement in error in outputs and inputs. Another problem is originated when a high number of inputs is considered: given enough inputs, all or most of the firms may be rates as efficient. In order to get reliable DEA results, Banker (1989) pointed out that the minimum number of units should be equal to or greater than three times the sum of the inputs and outputs.

3.2. Tobit regressions: "explaining" firms' technical efficiency

In order to examine IT's impact on the technical efficiency in the production process, we carry on the second stage of our study by regressing the scores of technical efficiency, derived from DEA in the first stage, against the respective IT variables.

Since technical efficiency scores by design are bounded to take values between 0 and 100, estimation by ordinary least squares would provide biased estimates. To avoid this problem, literature suggests estimating a Tobit model in which technical efficiency (TE) scores are transformed into technical inefficiency (TI) by the rule TI=100-TE. Within this framework, the Tobit regression model is formulated as following:

$$TI^*=X\beta+u$$

$$TI=0 \text{ if } TI^*\leq 0$$

$$TI=TI^* \text{ if } TI^*>0 \qquad (2)$$

where X is the vector of explanatory variables, β the vector of unknown parameters to be estimated and $u/X \sim N(0,\sigma^2)$.

4. Data

Data used in this study matches two national sources: the Survey on the Use of Information and Communication Technologies (ICT) in Enterprises, and the Annual Business Survey, both conducted by the Portuguese National Institute of Statistics. The former, which has been carried out since 2001, is part of the Eurostat initiative to collect comparable data for ICT adoption and use on a European scale; while the latter provides information on the main characteristics of Portuguese firms and their inputs and outputs (human resources, capital stock and added value, among others). The matching procedure is based on linking the ID numbers which identify firms in both surveys. Matching data for 2002's surveys gives a cross-section of 253 observations for small and medium size enterprises[3] (approximately 28% of firms are from manufacturing and

[3] The sample comprises firms with a number of workers between 10 and 250.

72% are in services), with complete information on both IT use and production function inputs and outputs.

With the described dataset, the following variables have been derived. In order to measure firm's technical efficiency we focus on the variables concerning outputs and inputs of production as follows: output (Y) is approximated by value-added, measured in euros at current prices and computed subtracting materials from the output; labour inputs (L) are represented by the total number of persons employed by the firm; capital inputs (K), in euros at current prices, are represented by the gross book values of all tangible and intangible assets of the firms as in Mairesse, Greenan and Topiol-Bensaid (2001). Then, to check the effect of IT over firm's efficiency we have considered two variables: computer capital (KC), in euros at current prices, represented the number of personal computers multiplied by their average value reported by IDC (2003) and computer usage (L_C) provided by the proportion of employees that work regularly with computers. As highlighted in section 2, the economic impacts of IT occur primarily when accompanied by a high-qualified labour force and by investments in complementary factors such as skills and organization practices. Therefore, the two following variables have been derived from the database: the quality of the workforce measured by workers educational level (L_S) and given by the proportion of workers with a university degree; and, outsourcing practices (Org) as a proxy for organizational (complementary) factors and given by a series of dummy variables that mirror the way firms solve their Information Technology challenges/problems

Table 1 shows the variables constructed with the dataset described above, as well as some descriptive statistics, by size and economic sector, since the sample is stratified according to these criteria.

Table 1. Data description (mean values for the year 2002)

Code	Variable	Description	Manufacturing		Services	
			Small	Medium	Small	Medium
Y	Value-added	Output - Materials (in € at current prices)	1890538.5	4060560	2007711.3	4284510.2
K	Capital stock	Gross book vale of tangible and intangible assets (in € at current prices)	9179418.6	19664770	9875129.1	9834182.5
L	Labor	Total number of employees	40	135	34	115
KC	Computer capital	Number of Computers X average value (in € at current prices)	20039.8	51933.7	34361.7	87598.2
Ls	Workers educational level	Percentage of workers with a university degree	9.4	7.3	15.3	15.7
Lc	Computer usage	Percentage of workers who regularly work with computers	33.8	32	66.6	57.5
Org	Outsourcing practices	Equal to one if firm solve its IT challenges using external services, and zero otherwise.	0.11	0.24	0.31	0.315
Number of firms			9	62	52	130

Note: the category "small" comprises firms with less than 50 workers; the category "medium" comprises firms with a number of workers between 50 and 249.

5. Results

5.1. DEA results on firms' technical efficiency

As noted in section 3.1, the BBC model assumes variable returns to scale, and computes the scores of technical and scale efficiency for each firm in the data set[4]. The averages and standard deviations from the mean of both types of efficiency are presented in Table 2. Also the numbers of firms with technical efficiency scores equal

[4] However, although efficiency measurement may be different assuming variable or constant returns to scale, empirically in our case we do not find significant differences. The correlation coefficient between both measurements is relatively high, 0.87, and significant at the 1% level.

to one (that is, those firms which are, comparatively speaking, the most efficient) are reported.

Table 2. Efficiency results from the BCC model

	Technical efficiency	Scale efficiency
Mean	0.284	0.868
Standard deviation	0.23	0.15
Number of efficient DMUs	10	10

The first point to note is that there is substantial technical inefficiency present in our sample. On average, Portuguese firms are operating at 28% efficiency. In other words, given inputs, SMEs could increase their output three-fold. A second point is the high mean value of scale efficiency, showing that most firms are close to operating on the right scale. Moreover, there are ten firms that get a score of 100% of efficiency.

It is important to take into account that the high-level of technical inefficiency observed could be due to both the DEA process and the structure of data. DEA is deterministic, and thus any noise in the data is treated as inefficiency; therefore, DEA results are highly sensitive to outliers. As Wang and Huang (2007) point out a DMU may be deemed as efficient in some circumstances merely by being different –in its output or output mix-from other units-. Nonetheless, this seems not to be our case, since the suppression of some units (suspected of being outliers) does not affect the average value, keeping at the level of 28%. Moreover, the low mean value of technical efficiency may reflect the heterogeneity of activities and firms, since our sample includes both manufacturing and services industries[5]. In fact, the estimated average efficiency for manufacturing industries under study is 20%, while this score increases to 31% in the case of service activities. The lower efficiency results for manufacturing

[5] Álvarez and Crespi (2003) also found disparities in efficiency scores among Chilean industrial activities, ranging from 91% in professional and scientific equipment to 34% in textiles activities (with average value for the full sample of 65%).

might be explained by the fact that the firms with a score of 100% all belong to the service sector. To take account of such heterogeneity, we will control for industry in the Tobit estimation.

5.2. Econometric results on correlation between IT and technical efficiency

Once obtained the technical efficiency scores, the Tobit model (2) is estimated considering two specifications. The first specification (A) checks the effect of IT on technical efficiency by including computer usage by workers as an explanatory variable. Meanwhile the second one (B) includes computer capital instead. Table 3 presents parameter estimates and the respective statistical tests for the whole sample of firms (columns 1-2), as well as by industry (columns 3-6).

When manufacturing and services firms are taken together in specification (A), all the coefficients have the expected sign and are statistically different from zero. In particular, results show that computer usage among workers has a positive significant effect on firms' technical efficiency. Since the dependent variable in the Tobit regressions is technical inefficiency, the negative coefficient of the variable computer usage means that the higher the proportion of workers using computers, the lower the technical inefficiency and, therefore, the higher efficiency. We also find that higher education has a positive influence on technical efficiency and is significant at the 1% level. Moreover, results point out that those firms where information technology challenges are solved using external resources are more efficient than the others (ceteris paribus). This is in accordance with information technology diffusion theory (Hong and Zhu, 2006), which sets that firms that count on partners or contractors for information technology implementation tasks are more likely to use these technologies intensively and efficiently.

Results for specification (B) suggest that computer capital stock, contrary to computer use, have a negative and statistically significant impact on technical efficiency, meaning that, ceteris paribus, firms with a high value of computer capital stock are, on average, less efficient than the others. Such results lead us to the following interesting point: for Portuguese SMEs' efficiency what matters is computer usage and workers' educational level and not the stock of computer capital.

By industry, we confirm the positive and significant impact of computer usage on the technical efficiency of manufacturing firms. Meanwhile in the case of services, the most important variables to explain technical efficiency are workers educational level as well as outsourcing practices.

Table 3. Tobit regressions

Dependent Variable: Technical Inefficiency						
	All sectors		Manufacturing		Services	
Independent Variables	(A)	(B)	(A)	(B)	(A)	(B)
Constant	80.7 (30.2)	32.55 (2.64)	53.98 (8.82)	114.4 (2.33)	79.52 (25.5)	21.1 (1.45)
Workers Educational Level (L_S)	-0.41 (-4.58)	-0.61 (-7.36)	0.083 (0.19)	-0.14 (-0.25)	-0.44 (-4.62)	-0.64 (-7.5)
Computer Usage (L_C) or Ln(KC)	-0.09 (-1.80)	4.27 (3.58)	-0.37 (-2.97)	-6.82 (-1.4)	-0.06 (-1.01)	5.39 (3.83)
Outsourcing (Org)	-1.88 (-1.93)	-1.34 (-0.46)	-10.6 (-1.34)	-10.34 (-1.23)	-1.97 (-2.48)	-1.99 (-0.56)
Manufacturing or Services (D)	5.8 (1.86)	7.88 (2.67)				
Log-Likelihood Value	-1109.5	-1104.04	-301.24	-303.06	-817.8	-805.4
$\hat{\sigma}$	420.7	453	894.3	932.6	421	572.8
Observations	253	253	71	71	182	182

5.3. Econometric results on correlation between IT and productivity

Our empirical analysis of Portuguese firms' productivity is based on an augmented Cobb-Douglas Production Function[6], where besides the classical labour and capital inputs, we also include the stock of computer capital, computer usage, workers level of education and outsourcing practices. Thus, we estimate the following two equations:

$$\ln(Y) = \alpha_0 + \alpha_1 \ln(K) + \alpha_2 \ln(L) + \alpha_3 L_C + \alpha_4 L_S + \alpha_5 Out + u \quad (3)$$

$$\ln(Y) = \alpha_0 + \alpha_1 \ln(K) + \alpha_2 \ln(L) + \alpha_3 \ln(KC) + \alpha_4 L_S + \alpha_5 Out + u \quad (4)$$

We assume that both models verify the classical linear regression hypothesis and the estimation procedure is, as usual, the Ordinary Least Squares Method (OLS). Table 4 presents the OLS results for the production function. First we estimate the model for all firms and then we desegregate by industry. The variable related to outsourcing partner usage is a dummy variable that is equal to one if firm solve IT challenges contracting resources from outside and zero otherwise.

[6] Given that we use cross-sectional data, our analysis is static and it is not possible to derive productivity from technical efficiency results.

Table 4.Production Function Estimation Results

Dependent Variable: ln(Y)						
	All sectors		Manufacturing		Services	
Independent Variables	(3)	(4)	(3)	(4)	(3)	(4)
Constant	6.38 (13.2)	6.62 (13.1)	5.75 (5.4)	3.74 (3.45)	6.35 (11.1)	6.77 (11.5)
Ln(L)	0.68 (10.9)	0.68 (10.1)	0.51 (4.1)	0.11 (0.77)	0.71 (9.7)	0.74 (9.5)
Ln(K)	0.32 (9.73)	0.33 (9.73)	0.4 (5.3)	0.36 (5.26)	0.31 (8.3)	0.33 (8.6)
Workers Educational Level (L_S)	0.014 (5.14)	0.018 (7.18)	-0.0006 (-0.06)	-0.021 (-1.85)	0.014 (4.8)	0.019 (7.16)
ln(KC) or Computer Usage (L_C)	0.0034 (2.11)	-0.02 (-0.66)	0.0057 (2.12)	0.46 (4.24)	0.003 (1.54)	-0.07 (-1.5)
Outsourcing	0.16 (1.92)	0.18 (2.08)	0.29 (1.92)	0.31 (2.24)	0.12 (1.23)	0.17 (1.64)
Manufacturing or Services (D)	-0,09 (-0.99)	-0.18 (-1.82)	-		-	
$\overline{R}^2$	62.1%	61.5%	57.5%	64.3%	62.4%	62.4%
Observations	253	253	71	71	182	182

Note: The quantities in parentheses below the estimates are the t statistics.

Some interesting conclusions come out from these results. Regarding specification (3), the coefficient for computer usage is positive and statistically significant at the 5% level, for the whole sample and for manufacturing firms, which means that, in our sample of SMEs, computer usage matters. The proxy variable for the human capital has, as expected, a statistically significant positive coefficient on firm's productivity; this means that, within this context, employees' skills are particular relevant, as suggested by Bresnahan, Brynjolfsson and Hitt (2002). We note that this effect is more pronounced for the service sector. Moreover, firms that count on partners to solve IT challenges are more productive, ceteris paribus; this effect is particularly

relevant for manufacturing firms. Concerning specification (4), we can conclude that for our sample of SMEs computer capital matters but only for manufacturing firms.

Overall and as expected, the results for the estimated effects of IT on efficiency and on productivity are very close. Computer Usage is relevant for both the productivity and effeciency of SMEs. However, some differences exist between economic sectors. In particular, we find that IT (either measured as computer stock or use) has stronger effects in the manufacturing sector. Shin (2006) reaches the same conclusion when analyzing the effects of the adoption of Enterprise Application Software on business performance in a sample of Korean SMEs.

6. Conclusion

Since the early 1990s, IT has become a priority for business management and strategy especially in SMEs. In this context, it is important to analyze whether such technologies have real effects on business performance and what effects, if any.

This paper analyzes the impacts of IT on SME's efficiency and productivity, two performance measures closely related. Compared to previous analysis in this field, the distinct feature of this paper is that we evaluate IT effects on firm performance by taking into account both IT stock and the intensity of IT use.

We note that there is a high level of technical inefficiency among Portuguese SMEs. In particular, they are operating at 28% efficiency, that is, given inputs they could increase their output three-fold.

Such inefficiency could be corrected or, at least, palliate by an effective use of IT. In this sense, results show that what really matters for SMEs is IT use and application; while IT stock plays a secondary role. In particular, IT use has a positive

effect on firms' technical efficiency. Thus, IT is a valid factor for firms to capture quickly and efficiently the information they need and to optimise their production processes, all these translating into productivity gains.

Moreover, the effects of IT are stronger in the manufacturing sector than in the service sector. In line with previous evidence, we find that the mere adoption of IT does not necessarily the successful performance of business organizations: our results show that adoption and investment in IT has to be accompanied by a high-qualified labour force and appropriate organisational practices.

The analysis presented in this paper can be extended in a main direction: the incorporation of the dynamics effects of IT stock and IT use on firm performance. We have run a static analysis due to the lack of data. If time series data were available, richer analysis could be done since the benefits from IT may take place after several years, when firms' complementary factors are conveniently adjusted.

REFERENCES

- Álvarez, R.; Crespi, G. (2003): "Determinants of technical efficiency in small firms", *Small Business Economics*, 233-244.
- Arvanitis, S. (2005): "Computerization, Workplace Organization, Skilled Labour and firm productivity: Evidence for the Swiss Business Sector", *Economics of Innovation and New Technology*, 14(4), pp. 225-249.
- Black, S.E.; Lynch, L. M (2001): "How to compete: the impact of workplace practices and information technology on productivity", *Review of Economic and Statistics*, 83 (3), pp. 434-445.

- Bresnahan, T.F.; Brynjolfsson, E.; Hitt, L.M. (2002): "Information Technology, Workplace Organization, and The Demand For Skilled Labor: Firm-Level Evidence", *The Quarterly Journal of Economics*, 117(1), pp. 339-376.
- Brynjolfsson, E.; Hitt, L.M. (1996): "Paradox Lost? Firm-Level Evidence on the Returns to Information Systems Spending*", Management Science*, 42(4), pp. 541-558.
- Brynjolfsson; E.; Hitt, L.M. (2003): "Computing Productivity: Firm-Level Evidence", Review of Economics and Statistics, 85(4), pp.793-808.
- Banker, R.D.; Charnes, A.; Cooper, W.W. (1984): "Some models for estimating technical and scale inefficiency in data envelopment analysis", *Management Science*, 30, pp. 1078–1092.
- Banker, R.D. (1989): "Econometric estimation and data envelopment analysis", Research in Governmental and Nonprofit Accounting (Ed.: J.L. Chan and J.M. Patton), 5, JAI Press, Greenwich, , pp. 231–243
- Charnes, A.; Cooper, W.W.; Rhodes E. (1978): "Measuring the efficiency of decision making units, *European Journal of Operational Research*, 2, pp. 429–444.
- Charnes, A.; Cooper, W.W.; Seiford, L.M. (1994): Data Envelopment Analysis: Theory, Methodology and Applications, Kluwer Academic Publishers.
- Draca, M.; Sadun, R.; Van Reenen, J. (2007): "Productivity and ICT: A review of evidence", *The Oxford Handbook of Information and Communication Technologies*, (Ed.: R. Mansell, C. Avgerou, D. Quah, R. Silverstone), Oxford University Press, 100-147.

- Dedrick, J.; Gurbaxani, V.; Kraemer, K.L. (2003): "Information technology and economic performance: A critical review of performance evidence", *ACM Computing Surveys*, 35(1), pp. 1-28.
- Dholakia, R.R.; Kshetri, N. (2004): "Factors Impacting the Adoption of the Internet among SMEs", *Small Business Economics*, 23, pp. 311–322,
- Farrell, M.J. (1957): "The measurement of productive efficiency", *Journal of the Royal Statistical Society*, Series A CXX, pp. 253–290.
- Hong, W., & Zhu, K. (2006) 'Migrating to internet-based e-commerce: Factors affecting e-commerce adoption and migration at the firm level', *Information and Management* 43: pp. 204-221.
- IDC (2003): *Mercado Nacional de PCs: Análise e Previsões, 2003 – 2008*, IDC. Lisbon.
- INE (2005): *Empresas em Portugal 2005*, Instituto Nacional de Estatística, Lisbon.
- Johnston, D.A.; Wade, M.; McClean, R. (2007): "Does e-Business Matter to SMEs? A Comparison of the Financial Impacts of Internet Business Solutions on European and North American SMEs", *Journal of Small Business Management,* 45(3), pp. 354-361.
- Lin, W.T.; Shao, B.B.M. (2000): "Examining the determinants of productive efficiency with IT as a production factor", *Journal of Computer Information Systems*, 41, pp.25–30.
- Lin, W.T.; Shao, B.B.M. (2006): "Assessing the input effect on productive efficiency in production systems: the value of information technology capital", *International Journal of Production Research*, 44 (9), pp.1799 – 1819.

- Lucchetti, R.; Sterlacchini, A. (2004): "The Adoption of ICT among SMEs: Evidence from an Italian Survey", *Small Business Economics***,** 23(2), 151–168.
- Mairesse J.; Greenan N.; Topiol-Bensaid A. (2001): "Information Technology and Research and Development Impacts on Productivity and Skills: Looking for Correlations on French Firm-Level Data", *NBER working paper 8075*, Cambridge, MA.
- Milana, C.; Zeli, A. (2004): "Productivity slowdown and the role of ICT in Italy: a firm-level analysis", The economic impact of ICT. Measurement, evidence and implications, OCDE, Paris, pp. 261-277.
- Morikawa, M. (2004): "Information Technology and the Performance of Japanese SMEs", *Small Business Economics*, 23, pp. 171–177.
- OECD (2002): "The impacts of electronic commerce on business: summary", OECD, Paris.
- OECD (2003): ICT and Economic Growth – Evidence from OECD Countries, Industries and Firms, OECD, Paris.
- OECD (2004): *ICT, e-business and SMEs*, OECD, Paris.
- OECD (2006): *Structural and Demographic Business Statistics database: 1996-2003*, OECD, Paris.
- Shao, B.B.M.; Lin, W.T. (2002): "Technical efficiency analysis of information technology investments: a two-stage empirical investigation", *Information & Management*, 39, 391-401.
- Shin, I. (2006): "Adoption of Enterprise Application software and firm performance", *Small Business Economics*, 241-256.

- Song, G.S.; Mueller-Falcke, D. (2006): "The Economic Effects of ICT at Firm-Levels*", Information and Communication Technologies for Development and Poverty Reduction. The Potential of Telecommunications* (Ed.: M. Torero and J. Von Braun), Johns Hopkins University Press, Baltimore, pp. 166-184.
- Sung, N. (2007): "Information technology, efficiency and productivity: evidence from Korean local governments", *Applied Economics*, 1 – 13.
- Wang, E.C; Huang, W. (2007): "Relative efficiency of R&D activities: A cross-country study accounting for environmental factors in the DEA approach", *Research Policy*, 36, pp. 260-273.
- Zwick, T. (2003): "The Impact of ICT Investment on Establishment Productivity", *National Institute Economic Review*, 184, pp. 99-110.

How does ICT enhance productivity? Evidence from latent retail technologies in Chile ⋆

March 2008

Gaaitzen J. de Vries *,[a], Michael Koetter [a]

[a] *University of Groningen, Faculty of Economics and Business, PO Box 800, 9700 AV Groningen, the Netherlands*

Abstract

It is widely known that the production technology of firms differs with the adoption of information and communication technology (ICT). Because ICT diffusion is incomplete, especially in developing countries, different groups of firms will have different production technologies. This needs to be accounted for when measuring productivity. In this paper, we estimate a latent class stochastic frontier model, which permits to test for the existence of multiple production technologies across firms and the associated implications for productivity measures. We use a unique data set of Chilean retailers, which includes detailed information on ICT adoption at the firm level. We identify four distinct groups for which ICT is an important determinant of the production technology. A higher use of ICT increases the probability of membership to a group with higher productivity.

Key words: ICT, Retail, Chile, Latent Class Stochastic Frontier Model
JEL: C23, D24, O33

1 Introduction

The crucial importance of information and communication technology (ICT) to conduct business and to streamline operations is virtually unequivocally

⋆ We would like to thank the National Statistical Office of Chile (INE), in particular Hernan Frigolett and Cristina Silva, for providing access to the firm level data. We thank Bart Los, Marcel Timmer, and seminar participants at the University of Groningen for helpful comments. Michael Koetter gratefully acknowledges support from the Netherland's Organization for Scientific Research NWO.

* Corresponding author
Email addresses: g.j.de.vries@rug.nl (Gaaitzen J. de Vries), m.koetter@rug.nl (Michael Koetter).

accepted (Brynjolfsson and Hitt, 2000; OECD, 2003; Bartel et al., 2007). Benefits from the use of ICT hold in particular for retailers (McKinsey, 2001). Investment in ICT can improve firm performance directly. For example, bar codes and scanners reduce checkout time and eliminate the need to manually price tag products thereby reducing labour costs.

Retailers can also benefit indirectly, for example from the use of computers for administration, inventory control, storage optimization, and pricing and promotion of products (McKinsey, 2001). Such ICT effects on firm performance may require substantial organizational changes, which in fact might indicate the use of a fundamentally different production technology. Such measures may potentially yield sustained improvements due to an improved matching of inventory to customer demand, more responsive price changes, more efficient use of shelf space, reduced inventory and fewer out-of-stock situations, the potential to evaluate and optimize advertising campaigns, and more efficient use of trucking and shipping (McGuckin et al., 2005).

Distinguishing ICT effects on firm productivity is complex and the few studies on retail firm-level productivity usually specify ICT proxies simply as an additional production factor (OECD, 2003; Broersma et al., 2003; Doms et al., 2004). Alternatively, we suggest in this paper a novel approach that treats observed indicators of ICT intensity as group membership probability determinants in different technology regimes. In that sense, we impose substantially less structure a priori on retail technologies and can test whether ICT use yields indirect benefits by increasing the odds for a firm to belong to a more productive technology regime.

Specifically, we examine the relation between ICT, productivity, and production technology of retailers in a developing country. Previous studies which examine the relation between ICT and productivity for retailers in developed countries assume a single production technology. However, ICT diffusion is incomplete, especially in rural areas of developing countries (Worldbank, 2008). So different groups of firms will have different production technologies depending on ICT adoption.[1] This needs to be accounted for when measuring productivity. Therefore, we augment a stochastic frontier model with a latent class structure as suggested by Greene (2005).

An alternative approach to account for cross-firm differences in production technology is to cluster firms, for example based on indicators of ICT adop-

[1] Incomplete ICT diffusion can prevail even if we assume that firms face similar ICT prices because of exogenous and endogenous factors which prevent the adoption of ICT. For example, other investments are necessary, for example cable networks, other infrastructure, or internet connections, to eliminate exogenous constraints on ICT adoption. Endogenous constraints related to ICT adoption include firm-level differences of technological literacy and skills to install and maintain ICT systems.

tion.[2] But clustering suffers from three important shortcomings. First, any a priori selection criteria is ultimately arbitrary. The common approach is to divide firms in a developing country by employment size (Tybout, 2000). However, some small firms use advanced technologies and should be compared with larger firms doing so as well, rather than with other small firms that use traditional technologies.[3] Second, the number of groups is unknown ex ante. Ideally the number of clusters should be borne out endogenously from the data by the extent of heterogeneity in production technology. Third, efficiency is a relative measure. Therefore, estimating stochastic frontiers after clustering the data set implies that relative efficiency scores cannot be compared across clusters.

In contrast to cluster analysis, the Latent Class Stochastic Frontier Model (LCSF) allows us to remain agnostic as to the number and composition of production technology regimes. In addition, cluster analysis splits a sample using the *value* of the separating variables, whereas the LCSF splits a sample according to the *effects* of the separating variables on the dependent variable (Corral and Álvarez, 2004).

We use a unique data set of approximately 1,100 Chilean retailers surveyed by the National Statistical Office's in its Encuesta Anual de Comercio for 2003 and 2004. This data set includes detailed information on ICT use for each firm, as well as balance sheet and supplementary economic information such as investment and the number of employees detailed by type. Our main result is the identification of four distinct technology regimes for which ICT determines group membership. A higher use of ICT increases the probability of membership to a group with higher productivity.

Our main finding is evidence in favour of four significantly different technology regimes among Chilean retailers. These groups differ with respect to the relative use of factors, as well as estimated efficiency and productivity. Firms in the most productive regime are approximately 40 percent more productive compared to firms in the least productive regime. We find that more intensive ICT use significantly increases the probability to belong to a more productive regime. Hence, our results are in line with other studies emphasizing the importance of ICT without imposing the rigid assumption that all firms operate identical technologies.

The remainder of this paper is structured as follows. In section 2 we present the method. Data and model specification are described in section 3. Results are discussed in section 4. Conclusions are presented in section 5.

[2] A related approach to cluster analysis is regression tree analysis, for example in Durlauf and Johnson (1995), who use values of initial GDP and literacy rates to identify multiple regimes across countries.

[3] In our data set of Chilean retailers, ICT adoption is not confined to larger firms.

2 Method

In this section we first introduce a fixed effects stochastic frontier model and note several limitations of this model. We then present the latent class stochastic frontier model and discuss its usefulness to account for the role of ICT to discern production technology regimes across Chilean retailers.

2.1 Fixed Effects Stochastic Frontier Model

Retailers use production factors, capital and labour, to sell goods and deliver services. Frontier analysis estimates the production technology of a firm by estimating the maximum possible output given a certain combination of inputs. [4] Deviations from optimal output measure Farrell (1957) type of inefficiency arising from the suboptimal use of input factors. A stochastic panel production frontier is written in logs as (Aigner et al., 1977):

$$y_{it} = \alpha_i + \beta' x_{it} + v_{it} - u_{it}. \tag{1}$$

where y is the log output of firm i at time t, and the matrix x_{it} includes the log of capital (K), high-skilled labour (Lhs), and low-skilled labour (Lls). To account partially for heterogenous ICT use in the production technologies across firms, we also specify firm-specific fixed effects α_i. As an improvement relative to conventional production function estimations in previous studies, we specify a composed error component accounting for technical efficiency, u. [5] It is measured as the ratio of observed output to the corresponding stochastic frontier output. The (exponent) value of technical efficiency ranges from 0 (fully inefficient) to 1 (fully efficient). For example, a firm exhibiting 20% inefficiency produced only 80% of its potential output had it employed its inputs efficiently. The random error term v accounts for statistical noise. To estimate equation (1) with maximum likelihood methods, we follow the convention in the stochastic frontier literature (Kumbhakar and Lovell, 2000) and assume that random error v_{it} is *iid* with $v_{it} \sim N(0, \sigma_v^2)$ and independent of the explanatory variables. The inefficiency term is assumed to be *iid* with $u_{it} \sim N|(0, \sigma_u^2)|$ and independent of v_{it}.

[4] The efficient production frontier can be obtained deterministically (Data envelopment analysis, DEA), which neglects possible measurement error. As an alternative, we therefore use here stochastic frontier analysis (SFA) to estimate the frontier instead (Coelli et al., 2005).

[5] This approach is still inflexible since factor elasticities are assumed to be constant across potentially different firms. We extend the model below for group-specific factor shares below.

Three issues deserve consideration. First, neglecting cross-firm heterogeneity may confound heterogeneity with inefficiency. Firm-specific effects α_i aim to capture heterogeneity. But in a disparate sample, fixed effects will capture much cross-firm heterogeneity as well as any inefficiency in the production process (Greene, 2005). Second, the production function literature has paid considerable attention to endogeneity issues regarding the input variables (see Olley and Pakes (1996) and Wooldridge (2005)). That is, inputs are correlated with errors due to unobserved factors, such as managerial quality. If problems with the omission of unobserved factors are not properly addressed, estimated coefficients will be biased. The specification of a stochastic frontier model allows for inefficiency, for example, due to poor management. It is therefore to a lesser extent subject to this concern. However, we do have several other endogeneity concerns, specifically those related to the correlation between ICT adoption and (capital) inputs, which we consider further below. Third, we abstain from specifying ICT as an additional production factor in equation (1). Instead, we hypothesize that factor elasticities might differ across firms and the production technology depends on ICT use. [6] Heterogeneity in production technology, however, is hard to define in terms of ICT use a priori, and it appears appropriate to model inefficiency and heterogeneity separately in the same model to segregate both effects. Therefore, we turn next to a new approach in productivity analysis, to account for heterogeneity in production technology.

2.2 Latent Class Stochastic Frontier Model

To model inefficiency and heterogeneity separately, we use a latent class stochastic frontier model proposed by Greene (2005). While latent class models are frequently used in mixture analysis (McLachlan and Peel, 2000) [7], the adaptation to frontier analysis is fairly recent. Greene (2003) segments different health care systems based on their orientation, for example, towards AIDS in developing African countries and cancer in developed OECD countries. Orea and Kumbhakar (2004) use the LCSF to study Spanish bank efficiency and find that banks can be grouped according to business scope and size. In this paper, we examine whether retailers can be grouped based on ICT use. Following Greene (2005), we write the latent class stochastic frontier model as:

$$y_{it} = \beta_j' x_{it} + v_{it|j} - u_{it|j}. \tag{2}$$

[6] One may argue, however, that ICT capital should be considered as a separate production factor as well and thus split up capital in Non-ICT capital and ICT capital. We agree but do not have detailed information on capital assets to split up physical capital.

[7] Mixture analysis estimates a "finite mixture" distribution, usually combined with Poisson regression analysis.

In contrast to the fixed effects stochastic frontier in equation (1), parameters differ across the latent classes $j = 1, ..., J$ and firm-specific effects α_i are dropped. We thus assume a latent sorting of retailers in the data set into J latent classes. Equation (2) is estimated using maximum likelihood methods. Maintaining the standard frontier assumption of a half normal distribution of the inefficiency term, the likelihood function is:

$$LF(i,t|j) = f(y_{it}|x_{it}, \beta_j, \sigma_j, \lambda_j) = \frac{\phi(\lambda_j \epsilon_{it|j})}{\phi(0)} \frac{1}{\sigma_j} \phi \frac{\epsilon_{it|j}}{\sigma_j}, \tag{3}$$

where $\epsilon_{it|j} = y_{it} - x'_{it}\beta_j$, $\lambda_j = \sigma_{uj}/\sigma_{vj}$, $\sigma_j = \sqrt{(\sigma^2_{uj} + \sigma^2_{vj})}$ and ϕ is the standard normal density. Conditional on the firm being in class j, the contribution of each firm to the likelihood function is:

$$LF(i|j) = \prod_{t=1}^{T} LF(i,t|j). \tag{4}$$

The unconditional likelihood for each firm is averaged over the latent classes using the prior probability as weights to membership in group j:

$$LF(i) = \sum_{j=1}^{J} P(i,j) LF(i|j) = \sum_{j=1}^{J} P(i,j) \prod_{t=1}^{T} LF(i,t|j). \tag{5}$$

In equation (5), the term $P(i,j)$ is the prior probability, which is attached to membership of firm i to class j. Firms reside in a class permanently. This prior probability therefore reflects the state of nature. The probability is specified for each firm if there are characteristics, z_i that sharpen the prior. A convenient parametrization of group membership is the multinomial logit form:

$$P(i,j) = \frac{\exp(z'_i \pi_j)}{\sum_{j=1}^{J} \exp(z'_i \pi_j)}, \pi_J = 0. \tag{6}$$

where, $j = J$ is the last group serving as the reference group and z_i are firm specific characteristics that determine class membership. Firm characteristics are assumed exogenous to input variables in the production function. [8]

In our specification, ICT determines class membership and we hypothesize that ICT is a determinant of different production technologies. However, if no firm characteristics are included in the estimation, the model still estimates latent classes (in that case, $P(i,j)$ would be a constant $P(j)$). Thus not only

[8] We address potential endogeneity in section 4.

firm characteristics, but the overall fit of the stochastic frontiers are used during the maximum likelihood procedure as well.

Summarizing, the LCSF estimates class-specific output coefficients β_j of the production factors capital, and high- and low-skilled labour, firm-specific efficiency u_{it}, and estimates whether ICT z_{it} affects the probability for firms to employ a different production technology. Firms belong to a latent class on the basis of probabilities from equation (6). It should be emphasized, that the efficiency of a firm is estimated relative to the frontier of its class. [9] Hence, productivity and efficiency should be carefully distinguished. The average productivity of retailers in the different classes can be directly compared and straightforwardly interpreted. However, efficiency of retailers is measured by the firms' position to its appropriate technology frontier, that is $v_{it|j}$.

The latent class model requires the number of groups J to be specified ex ante. In principle, the number of groups is only bounded by the number of cross-sectional units analysed. Since such a specification would suffer from over-specification problems, Greene (2005) suggests a top down approach based on likelihood ratio tests, because if there are J groups then estimates based on $J-1$ groups are inconsistent. We start by specifying five groups, compare it to a model with four groups $(J-1)$ and identify the number of groups based on likelihood ratio tests. In addition, we also perform Wald tests of the significance of differences between individual class parameters.

3 Data And Model Specification

3.1 Data

We apply the latent class stochastic frontier model to a short and largely balanced panel data set of retailers from the commercial survey (Encuesta Anual de Comercio, EAC) for 2003 and 2004. The commercial survey is conducted annually by the statistical office of Chile and covers a sample of approximately 1,100 retail firms. [10] Firms report in EAC: (a) balance sheet and income state-

[9] Group membership is based on the posterior probability. An alternative to calculate efficiency is to sum all posterior probabilities multiplied by the efficiency in using the technology of class j (Orea and Kumbhakar, 2004). The difference between both efficiency estimates is higher when the highest posterior probability is lower.

[10] The sample of firms in the commercial survey is defined from firms that are registered at Servicio de Impuestos Internos (www.cii.cl) (Declaración Annual de Impuestos a la Renta, formulario 22 y Declaración Mensual del IVA, formulario 29). The final sample is defined from firms with accumulated sales of 95 percent for the sector. This cut off at 95 percent is due to a large number of extremely small

ment information, such as cost, revenue, and profit information; (b) economic information beyond the balance sheet and income statement information, such as investment flows and the number of employees; (c) ICT information. We use two firm specific characteristics to capture ICT adoption among Chilean retailers. First, we use the number of computers per employee (where computers is the sum of desktop PC's, laptops, and servers). Second, we use detailed data on internet use. We create an ordinal value, labeled internet use, which ranges from 0 to 7 from the dummies of internet connection, intranet, extranet, e-mail address, website, purchases and or sales via the internet.

To measure retail output, several concepts can be used. In this paper, we use value added. The broadest output concept for distributive trade firms is sales. Sales are the number of goods sold multiplied by their respective price. [11] Using sales as the relevant output concept implies that both the product mix and the quantity of goods sold affect output. If the cost of goods sold is subtracted from sales, the resulting output concept is gross margin. [12] Thus, higher gross margins generally reflect higher value-added services. The gross margin output concept has several inherent difficulties. First, subtracting cost of goods sold from sales suggests that the costs of goods are separable from other costs the firm faces. Second, gross margins can be affected by volume discounts. Firms with market power might negotiate lower prices, increasing their gross margin. Third, volume measures of gross margin are difficult to measure since price data on cost of goods sold is needed. A third output concept is obtained by subtracting intermediate inputs from gross margin. This results in value added. Only labour and capital costs are included in the value added output concept. We use value added because it is common practice in national accounts. In addition, by using a value added output concept we are able to distinguish whether a retailer increased its value added output either by selling more or by reducing the costs of intermediate inputs. [13]

Firms report depreciation and investment in capital assets in EAC. Firms do not report gross capital assets. We assume that firms depreciate capital alike, and use reported depreciation as a proxy for the firms' capital stock. [14]

---

firms that are difficult to monitor and display huge instability over time. All large retailers are covered in the sample. Some firms that would significantly affect the precision of the aggregate variables are included as well (Inclusión Forzosa (IF) or forced inclusion). Other firms are sampled from the remaining population of firms.

[11] Sales with net inventory adjustment. Sales, wages, the cost of goods sold, and intermediate inputs for 2004 are deflated using the consumer price index.

[12] Preferably the gross margin output concept is extended by the provision of distribution services (Betancourt and Gautschi, 1993).

[13] For further discussion of the appropriate output concept for retailers, see McGuckin et al. (2005).

[14] Firms with a higher ICT capital stock are likely to have higher depreciation rates. However, ICT capital typically constitutes only a small share of the capital stock

The use of depreciation costs affects total factor productivity levels, but not the parameter estimates in the LCSF since the estimation procedure exploits variation in capital and not the level of capital.

Firms report the number of employees quarterly. We use the average annual employment as a measure of labour input. EAC distinguishes between various types of labour. We group these types into high-skilled labour (owners, executives, and managers), and low-skilled labour (family without fixed income, normal workers, temporary workers, and subcontracted workers). Descriptive statistics of the main variables are presented in table 1.

Our data set includes 920 retailers in 2003 and 906 in 2004. The data set is smaller than the original sample of approximately 1100 firms from the Encuesta Anual de Comercio. Some observations are lost because of missing information. Also, we correct for outliers. We trim the tails of the labour productivity distribution (VA/L), and the capital productivity distribution (VA/K).[15]

All quantity variables except computers per employee, are reported in logs in table 1. Output, measured by value added, increased from 2003 to 2004. Sales declined, but intermediate inputs declined even more, so the value added by retailers increased because of the more efficient use of inputs. In particular, table 1 shows indicators of ICT adoption by Chilean retailers. ICT diffusion is lower in Chile than in most developed OECD countries (OECD (2003)). For example, the proportion of businesses using the internet is above 80 percent in Japan, Australia, New Zealand, and Nordic countries. The share of businesses using the internet for purchases and sales ranges between 10 and 20 percent in these developed countries. Interestingly, ICT use by Chilean retailers is comparable with businesses in Greece. It should be noted, however, that our data set is not fully comparable since we only look at retailers, and we also leave out unregistered retailers which constitute a large share of the retail sector in Chile.

Most indicators in table 1 show an increase in ICT adoption by Chilean retailers. The number of computers per employee was 0.23 on average in 2003 and increased to 0.28 in 2004. In 2003, 49 percent of the retailers in our data set had an internet connection, 18 percent a website, and 3-4 percent sold or purchased goods via the internet. In 2004, the share of firms with an internet

---

and cross-firm differences in depreciation rates are therefore unlikely to be biased by this effect.

[15] We trim the 2.5 percent tails, which is somewhat higher than the common trimming of 1 percent tails. Our higher outlier correction is motivated by the likelihood of larger measurement error for a sample of services firms in a developing country. The main results in this paper are robust to the trimming of 1 percent tails, but there are differences at more detailed levels.

connection increased to 54 percent, and the share of firms with an e-mail address rose from 43 to 49 percent. The share of firms with a website, and the share of firms which sold or purchased goods via the internet hardly changed from 2003 to 2004. [16]

3.2 Model Specification

To estimate the production technology of retailers, we specify a translog functional form:

$$\begin{aligned}\ln Y_{it|j} = {} & \alpha_j + \beta_{1j} \ln K_{it} + \beta_{2j} \ln Lhs_{it} + \beta_{3j} \ln Lls_{it} + \frac{1}{2}\beta_{4j} \ln K_{it}^2 \\ & + \frac{1}{2}\beta_{5j} \ln Lhs_{it}^2 + \frac{1}{2}\beta_{6j} \ln Lls_{it}^2 + \beta_{7j} \ln K_{it}. \ln Lhs_{it} \\ & + \beta_{8j} \ln K_{it}. \ln Lls_{it} + \beta_{9j} \ln Lhs_{it}. \ln Lls_{it} + v_{it|j} - u_{it|j}\end{aligned} \tag{7}$$

In equation (7), subscripts i, t, and j refer to firm, time and class respectively. Y, K, Lhs, and Lls denote output, capital, high-skilled labour, and low-skilled labour, respectively. As separating variables in the identification of latent classes we use our two proxies of ICT: the number of computers per employee (CpE) and the intensity of internet usage. The intensity of internet usage is measured by the sum of dummies for internet connection, intranet, extranet, e-mail address, website, and purchases and sales via the internet. This value therefore ranges from 0 to 7. Latent class probabilities are written as:

$$P(i,j) = \frac{exp(\pi_0 \ + \pi_{1j} \ CpE_i + \pi_{2j} \ InternetUse_i)}{\sum_{j=1}^{J} exp(\pi_0 \ + \pi_{1j} \ CpE_i + \pi_{2j} \ InternetUse_i)}, \pi_J = 0. \tag{8}$$

In equation (8), the last class serves as the reference group. No time element is included in equation (8), so ICT characteristics for 2003 are used to determine class membership, and firms remain in a class throughout the period analysed. Since our data set covers two years only, not allowing transitions between production technology regimes is not a major concern.

We are, however, concerned about endogeneity in our model specification (7 and 8). In particular, firm characteristics related to ICT adoption might be correlated with inputs. For example, ICT might be correlated with high-skilled

[16] Unregistered firms are not sampled by the Encuesta Anual de Comercio. Since unregistered firms use less ICT, our data set overestimates ICT adoption by Chilean retailers.

labour. Ideally, we use ICT prices as an instrumental variable to address this endogeneity concern. Unfortunately, detailed ICT price data are not available. We consider the robustness of our results by using lagged ICT adoption and the change in ICT adoption. Both lagged ICT and changes in ICT adoption are likely to be less correlated with current inputs.

4 Results

4.1 Stochastic Frontier Estimation

We first estimate a standard fixed effects panel frontier model as in equation (1) on approximately 900 retail firms operating in 2003 and 2004. Results are shown in the first column of table 2. Output elasticities of capital, and high- and low-skilled labour are significant at the 5 percent level. Individual parameter estimates of λ and σ show that inefficiency prevails. Wald tests confirm that both inefficiency terms are individually and jointly significant. Hence, a stochastic frontier specification which permits inefficiency in the production process is the appropriate choice. In addition, a Wald test of the additional input coefficients from the translog functional form supports the specification of the translog as opposed to the Cobb Douglas functional form. [17]

Next, we extend the pooled stochastic frontier model to a model which specifically allows for heterogeneity in production technology due to differences in the adoption of ICT by Chilean retailers. We estimate a latent class stochastic frontier model to test if different technology regimes prevail. In this model we treat firm heterogeneity as being generated by a discrete distribution and specify the number of classes a priori.

We choose the number of classes using a 'top-down' approach, where the model with the largest log-likelihood ratio is preferred (Greene, 2005). We first specify five classes and examine changes in the likelihood ratio when reducing the number of classes. According likelihood values are shown in table 3.

The likelihood ratio is lowest when a model with four latent classes is estimated. Furthermore, if we estimate a model with five latent classes, individual parameters from one group are not significantly different from zero and we cannot reject the hypothesis that the joint coefficients of this group are equal to zero. This finding provides additional support for a model specification with four latent classes (so that $j = 4$).

[17] The P-value for the Wald test of no inefficiency is 0.00, and the P-value for the additional input coefficients from the translog functional form is 0.00 as well.

4.2 ICT as Technology Regime Determinant

Estimation results of the LCSF with four classes are shown in table 2. Parameter estimates of the output elasticities with respect to capital, high- and low-skilled labour are shown for each class. Note that each regime-specific vector of production technology parameters is estimated simultaneously. Some direct and interacted parameter estimates are negative, indicating decreasing returns to scale for individual input factors in some groups. Scale economies at the firm level equal the sum of these partial derivatives per input with respect to output. For each technology regime of retailers these are larger than unity, indicating the presence of increasing returns to scale at the firm level. Individual parameter estimates for σ and λ suggest that retailers in all classes are fairly efficiently operating their appropriate technology. In the bottom row, $P(q)$ indicates group membership probabilities conditional on ICT adoption. Approximately 21 percent of retailers in our sample belong to the first class. This compares with 16 percent in the second class, 48 percent in the third class, and 15 percent in the fourth class.

Of particular interest are the ICT coefficients in the latent class probability functions. For all classes except the third class ICT coefficients are statistically significant at the 1 percent level. Therefore, ICT provides useful information in classifying the sample. Retailers do not share a common technology and ICT significantly predicts production technology regime membership. For the first class, we find a significant positive sign (1 percent level) for the coefficient of Computers per Employee (CpE) and also for our indicator of $InternetUse$. This implies that higher ICT adoption increases the probability for a retailer to belong to the first technology regime relative to belonging to the control group, group four in our case. The results for the second class highlight that our proxies of ICT used here measure different aspects. While more computers per employee (CpE) increase regime membership likelihood, higher $InternetUse$ reduces the odds to belong to this regime. Membership in the third class, in turn, does not depend significantly on ICT indicators, potentially indicating that ICT is not a crucial component of these retailers business model.

Before investigating the characteristics of the four different regimes in more detail, note that the parameter estimates of the fixed effects frontier lie within the range of parameters from the latent classes (see table 2). Wald tests indicated already that parameters of the latent classes are significantly different. This result shows that the assumption of a single production function, whether a frontier or OLS, is an assumption which fails to adequately capture systematically different production technologies.

Descriptive statistics of the latent classes are presented in table 4. We show relative total factor productivity (TFP) and labour productivity based on esti-

mated parameters. We show relative TFP levels, since the use of depreciation costs as a proxy for capital does not affect capital and labour coefficients, but it does affects TFP levels. Therefore, we set TFP at 100 for class 4, which serves as the benchmark. We find a higher TFP in the first, second, and third class. Labour productivity largely mimics the TFP pattern, since labour productivity is higher in the first and third class. Labour productivity, however, is lower in the second class as compared with the fourth class. This is due to higher capital intensity of firms in class 4.

Productivity is higher in classes with more intensive ICT users. In particular, the first class uses ICT most intensively and is also the most (both labour and TFP) productive regime. In 2003, two computers were available for every five employees in retailers in the first class. This compares with one computer for every 10 employees in the fourth class. Indicators of internet use confirm differences as well. For example, the share of retailers in the first class with an internet connection was 68 percent in 2003. This compares with 37 percent in the fourth class. While 34 percent of retailers in the first class had a website, this is only 11 percent in the fourth class. Productivity is higher in the first class. For labour productivity it is 1.3 log points higher, and relative TFP levels indicate that retailers in the first class are 40 percent more productive on average.

Productivity and efficiency in table 4 should be carefully distinguished. The average productivity of retailers in the different classes can be directly compared and straightforwardly interpreted. However, efficiency of retailers is measured by the firms' position to its respective, group-specific technology frontier, that is $u_{it|j}$. For example, most retailers in the third class are close to their technology frontier. Many retailers in the second class are far from their frontier. So ample scope exists in the second class to increase productivity by reducing inefficiency and thereby moving closer to their appropriate technology frontier.

Four main differences across Chilean retailers emerge from the distinction of technology regimes by the LCSF model. First, firms in the first class are largest on average. These firms have the highest number of unskilled employees, which are probably hired to stock shelves and check out customers. These firms make more use of the more "advanced" ICT options, such as realising sales and purchases via the internet. Second, firms in the second class are smallest on average. These firms have the highest high/low skilled ratio, which in combination with the negative parameter coefficient for high skilled labour in the second class (see table 2), suggests adding more managers is not an output enhancing strategy for them. Firms in the second class show substantial variation in efficiency as well. Third, most firms are in the third class. Factor inputs of these firms are comparable with firms in the fourth class except for the larger number of unskilled workers. Firms in this class operate their production technology efficiently. And ICT adoption is somewhat higher than

firms in the fourth class. Finally, firms in the fourth class are least productive when measured by TFP, and average ICT adoption is lower compared to other classes. In addition, these firms have substantial scope to improve efficiency.

4.3 Robustness Checks

Based on the estimation of a LCSF with four classes, we find that retailers do not share a common technology and ICT significantly increases the probability of more productive production technology regime membership. Here, we examine the robustness of these results.

First, although our statistical tests indicated a clear preference for a model with four classes, the role of ICT in shaping heterogeneity might depend upon the number of latent classes specified a priori. Yet, if we estimate LCSF with five classes we find similar results. While ICT still controls for existing heterogeneity in most of the different technology regimes, we cannot discern the additional group significantly based on ICT effects alone. For this class all parameters and also the joint significance is not significantly different from zero. Hence, a fifth group is not sufficiently different in terms of ICT use to warrant an additional technology regime with significantly different production characteristics. Further, if we estimate a LCSF with three latent classes, we also find a significant effect of ICT to predict production technology group membership.

Second, our results might depend upon the functional form. We therefore specified a standard Cobb Douglas production function, too. In a Cobb Douglas framework, value added depends upon the inputs (capital, and high- and low-skilled labour) in a linear fashion. Although statistical tests reject a Cobb Douglas specification, we examine these results to reconcile our findings with the extant literature that uses explicitly or implicitly such a functional form. A LCSF with four classes and a Cobb Douglas functional form also yields that ICT significantly controls for heterogeneity of production technologies. Intensive ICT use increases the probability of belonging to a more productive technology regime, which corroborates our previous findings.

Third, we are concerned about endogeneity. In particular, the adoption of ICT might be correlated with inputs. For example, ICT adoption might be related to the educational level of employees because of skill-biased technology adoption. ICT adoption might also be correlated with capital inputs.[18] We aimed to control for potential endogeneity in two ways. First, we used the change in ICT adoption. This effectively halved our sample and changed the results in

[18] We find a positive correlation between ICT, capital, and labour skills in our dataset.

various ways. Likelihood ratio's now indicate a preference for a model with five classes. In addition, ICT does not significantly affect group membership for two classes. But ICT does significantly affect group membership for the other two classes and these two classes are more productive than the control group. Second, we used one-year lagged ICT to control for endogeneity. This reduces our sample once again, since we lose observations for 2003. Results, however, change only slightly. Likelihood ratio's again indicate a preference for a model with five classes. But we still find that ICT is a significant determinant of technology regime membership. And ICT adoption increases the probability of belonging to a more productive technology regime.

Fourth, the results might be sensitive to the ICT indicator used. We therefore estimated the LCSF with four classes using either computers per employee or internet use. Results are comparable if we only use computers per employee as an indicator of ICT adoption. Computers per employee significantly affect group membership except for the third class. If we use internet use to examine the role of ICT as a technology regime determinant, we find that internet use is a significant determinant for membership to the first group only.

Therefore, our main results are robust. Chilean retailers have significantly different production technologies and membership probabilities in these different regimes depends on ICT use.

5 Conclusion

In this paper we take a novel approach to examine the relation between ICT and the productivity of retail firms in a less developed country. Methodologically, we seek to advance by estimating in a single stage a latent class stochastic frontier model as to obtain class-specific production frontier parameters, firm-specific inefficiency, and the likelihood for a firm to belong to a latent technology regime explained by ICT adoption. Thereby, we avoid a priori choices on firm grouping as in cluster analysis as well as the impossibility to conduct relative comparisons from separate frontier estimations in the previous productivity literature. To this end, we use a unique and largely balanced panel data set provided by the Chilean statistical office which includes detailed firm-level data on financial accounts, ICT use, and further economic information for 2003 and 2004. We find three main results.

First, many Chilean retailers employ production factors in sub-optimal proportions. The preferred specification is therefore a neo-classical production frontier that allows for technical inefficiency. Furthermore, a translog flexible form is preferred to the more frequently used Cobb-Douglas specification.

Second, we find strong evidence in favour of multiple production technologies. We identify four distinct groups. These groups differ in terms of productivity and ICT adoption. Firms also differ in their ability to exploit their appropriate technology's production possibilities. In particular, we find more intensive ICT users and a 40 percent higher TFP level in the first group than the fourth (control) group.

Third, firm-specific regime membership probabilities depend significantly on the adoption of information and communication technology (ICT). Consistent with previous evidence for other industries and for developed countries, we find that ICT is positively related to productivity, since higher ICT adoption increases the likelihood for a firm to belong to a more productive technology regime.

References

Aigner, D., C. A. K. Lovell, and P. Schmidt (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics 6*(1), 21–37.

Bartel, A., C. Ichniowski, and K. Shaw (2007). How Does Information Technology Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement, and Worker Skills. *Quarterly Journal of Economics 72*.

Betancourt, R. and D. Gautschi (1993). The Outputs of Retail Activities: Concepts, Measurement, And Evidence From U.S. Census Data. *Review of Economics and Statistics 75*.

Broersma, L., R. McGuckin, and M. P. Timmer (2003). The Impact of Computers on Productivity in the Trade Sector: Explorations with Dutch Microdata. *De Economist 151*, 53–79.

Brynjolfsson, E. and L. M. Hitt (2000). Beyond Computation: Information Technology, Organizational Transformation and Business Performance. *The Journal of Economic Perspectives 14*(4), 23–48.

Coelli, T., D. P. Rao, and G. E. Battese (2005). *An Introduction to Efficiency Analysis* (2 ed.). New York: Springer.

Corral, J. and A. Álvarez (2004). Estimation of Different Technologies Using a Latent Class Model. *Efficiency Series Paper Universidad de Oviedo 7*.

Doms, M., R. Jarmin, and S. Klimeck (2004). Information Technology Investment and Firm Performance in U.S. Retail Trade. *Economics of Innovation and New Technology 13*(7), 595–613.

Durlauf, S. N. and P. A. Johnson (1995). Multiple Regimes and Cross-Country Growth Behaviour. *Journal of Applied Econometrics 10*.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society*.

Greene, W. (2003). Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems. *Mimeo, New York University*.

Greene, W. (2005). Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model. *Journal of Econometrics 126*, 269–303.

Kumbhakar, S. C. and C. A. K. Lovell (2000). *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press.

McGuckin, R., B. van Ark, and M. Spiegelman (2005). The Retail Revolution: Can Europe Match U.S. Productivity Performance? Technical report, The Conference Board.

McKinsey (2001). US Productivity Growth 1995-2000: Understanding the Contribution of Information and Technology Relative to Other Factors. Technical report, McKinsey Global Institute, London.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley Sons.

OECD (2003). *ICT and Economic Growth: Evidence From OECD Countries, Industries, and Firms*. Paris: OECD.

Olley, S. and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica 64*, 1263–1298.

Orea, L. and S. Kumbhakar (2004). Efficiency Measurement Using a Latent Class Stochastic Frontier Model. *Empirical Economics 29*, 169–183.

Tybout, J. (2000). Manufacturing Firms in Developing Countries: How Well Do They Do And Why? *Journal of Economic Literature 38*, 11–44.

Wooldridge, J. (2005). On Estimating Firm-level Production Functions Using Proxy Variables to Control for Unobservables. *Mimeo, Michigan State University*.

Worldbank (2008). *Global Economic Prospects 2008. Technology Diffusion in the Developing World*. World Bank.

Table 1
Descriptive statistics

	2003		2004	
	mean		mean	
lnSales	13.00		12.99	
	2.00		*2.07*	
lnCost of goods sold	12.58		12.56	
	2.19		*2.25*	
lnValue added	10.86		10.93	
	1.98		*2.02*	
lnK	7.88		7.82	
	2.29		*2.43*	
lnLhs	1.00		0.96	
	0.57		*0.55*	
lnLls	2.60		2.59	
	1.67		*1.72*	
Computers per employee	0.23		0.28	
	0.43		*0.50*	
Number and share of firms with:	obs.	share	obs.	share
Internet connection	449	49%	486	54%
Intranet	119	13%	172	19%
Extranet	56	6%	53	6%
E-mail address	399	43%	448	49%
Website	161	18%	162	18%
Purchases via internet	35	4%	29	3%
Sales via internet	32	3%	27	3%
Observations	920		906	

Note: Values of sales, cost of goods sold, value added, capital, high- and low-skilled labour in 2004 are deflated. Standard deviations are reported in italics.

Table 2
Frontier analysis

	FESF	LCSF			
Production frontier		Class 1	Class 2	Class 3	Class 4
lnK	-0.25	-0.19	-0.16	0.07	-0.93
	-12.06	*-2.16*	*-0.49*	*1.78*	*-10.20*
lnLhs	0.41	0.42	-3.32	0.56	1.46
	5.49	*2.00*	*-1.86*	*3.90*	*3.89*
lnLls	1.31	0.92	2.28	0.92	2.55
	47.60	*8.83*	*6.59*	*18.91*	*17.75*
$\frac{1}{2}$lnK^2	0.07	0.04	0.01	0.01	0.29
	14.31	*2.35*	*0.18*	*1.23*	*21.21*
$\frac{1}{2}$lnLhs^2	0.01	-0.06	-0.31	0.17	1.96
	0.15	*-0.63*	*-0.15*	*2.40*	*6.01*
$\frac{1}{2}$lnLls^2	0.01	-0.06	0.27	-0.02	0.26
	0.69	*-1.55*	*2.27*	*-1.83*	*6.62*
lnK×lnLhs	0.03	-0.02	0.76	0.00	-0.47
	2.65	*-0.54*	*2.48*	*-0.15*	*-6.73*
lnK×lnLls	-0.05	0.00	-0.17	0.01	-0.32
	-8.77	*0.08*	*-3.55*	*0.87*	*-14.38*
lnLhs×lnLls	-0.11	0.00	-0.77	-0.14	0.20
	-6.94	*-0.09*	*-2.30*	*-4.45*	*2.18*
σ	1.23	0.34	1.98	0.28	0.64
	130.52	*4.25*	*26.51*	*0.14*	*15.47*
λ	2.59	0.65	17.34	0.00	36.05
	27.05	*0.73*	*1.38*	*0.00*	*0.86*
Computers per Employee	-	2.13	2.59	1.07	0.00
		2.95	*3.43*	*1.46*	
Internet Use	-	0.33	-0.25	0.08	0.00
		3.39	*-1.75*	*0.83*	
P(q)	-	0.21	0.16	0.48	0.15

Notes: FESF is fixed effects stochastic frontier. LCSF is latent class stochastic frontier. The number of observations is 1,826. Log likelihood ratio fixed effects stochastic frontier is -1415.07. Log likelihood ratio latent class stochastic frontier is -1544.39. $\lambda_j = \sigma_{uj}/\sigma_{vj}$, and $\sigma_j = \sqrt{(\sigma_{uj}^2 + \sigma_{vj}^2)}$, where σ_{uj} is the standard error of the inefficiency term and σ_{vj} the standard error of the random error term. P(q) refers to the group membership probabilities conditional on ICT use. β/s.e. are in italics below. Intercepts are not reported.

Table 3
Selection statistics

No. of classes	One	Two	Three	Four	Five
Log likelihood	-2044.84	-1800.62	-1614.02	-1544.39	-1567.93

Table 4
Descriptive statistics of latent classes

	Class 1		Class 2		Class 3		Class 4	
	2003	2004	2003	2004	2003	2004	2003	2004
Relative TFP	137.6	138.6	134.7	136.1	139.5	140.4	100.0	100.0
lnLabour productivity	9.0	9.1	7.1	7.3	8.3	8.4	7.7	7.8
	0.4	*0.4*	*1.3*	*1.3*	*0.3*	*0.4*	*0.4*	*0.6*
lnK	8.8	8.7	6.7	6.7	7.9	7.8	7.6	7.6
	2.6	*2.7*	*2.0*	*2.4*	*2.1*	*2.3*	*2.1*	*2.3*
lnLhs	1.2	1.2	0.9	0.8	1.0	0.9	0.9	0.9
	0.8	*0.8*	*0.3*	*0.3*	*0.5*	*0.5*	*0.4*	*0.4*
lnLls	3.2	3.2	1.7	1.6	2.7	2.7	2.3	2.3
	2.0	*2.0*	*1.4*	*1.5*	*1.5*	*1.5*	*1.6*	*1.7*
Computers per employee	0.4	0.5	0.3	0.5	0.2	0.2	0.1	0.1
	0.6	*0.7*	*0.7*	*0.8*	*0.2*	*0.3*	*0.2*	*0.2*
Internet use	2.1	2.5	1.0	1.0	1.2	1.4	1.0	1.1
	1.8	*1.7*	*1.4*	*1.3*	*1.5*	*1.5*	*1.5*	*1.6*
Share of firms with:								
Internet connection	68%	80%	40%	42%	47%	50%	37%	39%
Intranet	22%	37%	7%	10%	11%	15%	11%	15%
Extranet	14%	15%	2%	2%	5%	3%	4%	5%
E-mail adress	64%	75%	34%	38%	40%	46%	32%	35%
Website	34%	38%	9%	7%	15%	15%	11%	11%
Purchases via internet	5%	5%	3%	1%	4%	3%	2%	5%
Sales via internet	7%	6%	3%	1%	3%	3%	1%	1%
Technical efficiency	93%	93%	32%	36%	100%	100%	64%	65%
	2%	*2%*	*27%*	*27%*	*0%*	*0%*	*19%*	*20%*
Observations	197	191	117	108	455	457	151	150

Note: TFP levels are relative to class 4, which is set at 100. Labour productivity is value added divided by the sum of high- and low-skilled labour. Standard deviations are in italics.

Electronic intermediation and two-sided markets: what happens when sellers and buyers can switch?*

Pierre Gazé† and Anne-Gaël Vaubourg ‡

Abstract

In this paper, we define electronic platforms as two-sided markets in which both sides of the market can easily switch. To account for this specificity, we consider two platforms in a duopoly through which sellers and buyers can match during two successive sessions. Between the two sessions, some sellers become buyers and *vice-versa*. We show that equilibrium participation fees can be written as the sum of two terms. The first one is the equilibrium price without mobility. The second one can be interpreted in terms of "rewards" and "penalties" relatively to prices without mobility. As rewards do not perfectly compensate penalties, equilibrium prices can be higher or lower than prices without mobility. We also demonstrate that the platforms' profit increases with global mobility. It also increases with relative mobility provided this mobility is large enough.

JEL Codes: L11, L13, L86

Key words: Electronic intermediation, two-sided markets, externalities, mobility

*This work has benefited from the support of the CNRS under the program "ATIP Jeunes Chercheurs".

†LEO-Université d'Orléans, e-mail: Pierre.Gaze@univ-orleans.fr

‡LEO-Université d'Orléans, e-mail: Anne-Gael.Vaubourg@univ-orleans.fr, corresponding author

1 Introduction

The expression "new economy" has emerged at the end of the 1990s. It refers to a new hypothetic macroeconomic regime characterized by low inflation and, thanks to Internet, high growth rates. Since the explosion of the Internet bubble this expression have tended to disappear. But some of the initial interrogations remain. Does Internet lead to a specific economy? What is really new in electronic intermediation? Our paper adresses this issue to a two-sided market setting. The main idea of this work is that electronic intermediation gives birth to a new class of two-sided markets in which agents can easily switch from one side to another.

In two-sided markets (TSM), at least two groups of agents interact through an intermediary called a platform. Each group gives value to the participation of the other group. For this reason two-sided markets are characterized by a specific class of network externalities. Rochet & Tirole (2003) and Evans (2003) review many industries that exhibit such a feature. Videogame platforms, credit card payment systems and dating agencies provide well-known examples of TSM. An abundant and fruitful theoretical literature, surveyed by Rochet & Tirole (2006) and Roson (2005), has developed to explore the fundamental economic principles of these peculiar markets. Credit cards payment systems have been analysed in many papers (Rochet & Tirole (2002), Guthrie & Wright (2003) and Chakravorti & Roson (2006)). These markets are characterized by transaction fees, i.e. fees that are paid (by the sellers to the issuer of the card) when transactions occur. Gabszewicz, Laussel & Sonnac (2001), Gabszewicz, Laussel & Sonnac (2004), Anderson & Gabszewicz (2006), Ferrando, Gabszewicz, Laussel & Sonnac (2004) and Kaiser & Wright (2006) devote a special attention to another category of TSM: the media industry. When they buy a part of newspapers' or tv-channels support, advertisers pay a fee even if their insert is not seen by readers or viewers. These so-called participation (or registration) fees, which do not depend on whether a transaction has occured, prevail in other TSM such as nightclubs for example. Analysed by Armstrong (2006) in a canonic duopoly model with horizontally differentiated platforms, they are commonly observed when transactions are impossible or costly to monitor.

The Internet network gives an opportunity for several existing two-sided markets to be developed or reorganized. It is the case for auctions, traveling services or media industries. Some activities even owe their existence to the Internet network. Massively Multiplayer Online Role-Playing Games (MMORPG), the virtual purse or some types of electronic money belong to this category. Sometimes, it is the Internet network itself that appears as a TSM. For example, the Internet Service Provider

acts as a platform enabling connexions and transactions between web sites and consumers. In both cases (creating or developing preexisting activities), Internet is a new way of making transactions. Virtual worlds and massively multiplayer on-line virtual universes are quite anecdotical in terms of economic weight. That is not the case for on-line markets on the whole. They are clearly attractive to consumers in terms of selection, availability and prices, compared to their physical counterparts, thereby explaining their rapid growth[1]. On-line retail sales in the United-States reached $114 billion in 2006, up from $93 billion in 2005 (an increase of 22.7%). In the US, e-commerce sales in the first quarter of 2007 accounted for 3.2 percent of total retail sales (source: Monthly retail Trade Survey). This trend is similar in Europe with 42 billion euros being spent in 2006 in the United Kingdom, 22 billion euros in Germany and 12 billion euros in France (source: Forrester Research, Inc.). So there is much at stake in better understanding electronic intermediation and in determining to what extent this new form of intermediation generates new economic rules.

The specificities of Internet TSM have been examined by Caillaud & Jullien (2001) and Caillaud & Jullien (2003). They consider that electronic intermediaries are able to monitor transactions, which allows them to charge both participation and transaction fees to consumers. They also argue that Internet platforms give the agents the opportunity to make multi-homing, i.e. to register to several platforms. Caillaud & Jullien (2001) show that in a duopoly with registration transactions and single-homing, there exists an asymmetric equilibrium where a dominant firm captures the whole market and earns positif profit. When transaction fees are added to participation fees, the profit becomes zero. This findings also holds in a symmetric equilibrium (Caillaud & Jullien (2003)). When the single-homing assumption is alleviated, the dominant firm's profit becomes zero and the two sides of the market are offered a free access to the platform. In the symmetric equilibrium exhibited by Caillaud & Jullien (2003), the result is quite different: both platforms earn positive profit and only one group of agents benefits from free participation.

From our viewpoint, the two specificities pointed out by Caillaud & Jullien (2001) and Caillaud & Jullien (2003) have to be qualified. On the one hand, as underlined by Rochet & Tirole (2006), the charging of transaction fees is undermined by the possibility to bypass electronic monitoring and to materialize transactions off-line. On the other hand, although Internet facilitates the membership to several networks, multi-homing is also possible in many non-virtual TSM. This paper proposes another criterion that may better discriminate between one-line and off-line

[1]For a theoretical analysis of the diffusion of e-commerce, see Dinlersoz & Pereira (2007).

TSM.

The main idea of our paper is that electronic intermediaries are platforms in which buyers can easily become sellers and *vice-versa*. The eBay website auctions is characterized by cross-group network effects between buyers and sellers. Thanks to electronic intermediation, a buyer in a matching session can become a seller for the next session. Switching is easy, reversible and costless. Agents enjoy this switching possibility that allows them to be seller or buyer according to their needs and desires without any (or with light) institutional constraints. The most well-known electronic payment "Paypal" also has this peculiar design. PayPal is a person-to-person on-line payment instrument designed for any type of monetary transfer such as auctions (eBay), gifts, etc. between PayPal users. PayPal offers to one group of agents the possibility to send funds while it allows another group of agents to receive them. Contrary to a credit card payment system, an agent is not identified as a merchant or as a customer once and for all. An agent can belong to the group of payees for a session and move to the group of payers for the next session. What is very different from traditional on-line payment systems is that you do not need any complex hardware or software device to receive funds: electronic mail is enough. Some identical characteristics prevail in the Internet-based virtual world "Second Life". Each gamer can buy or sell items (virtual or real commodities or services) to improve his satisfaction. We can consider that the group of sellers and the group of buyers are in a perpetual reformation. In the Massively Multiplayer Online Role-Playing Games "Entropia Universe" gamers use the private money called Project Entropia Dollars (PED), that can be redeemed into real US dollars. Each gamer can spend or earn PED according to game circumstances.

In this paper, we conceive a model to analyze these specific classes of TSM. Extending Armstrong (2006)'s framework, we examine what happens when agents can migrate from one group to another by investigating how mobility affects platforms' equilibrium prices and profits. The remainder of the paper is organized as follows. In Section 2, we present the assumptions of the model. The equilibrium is analysed in Section 3. Section 4 concludes the article.

2 The model

2.1 Assumptions

Following Armstrong (2006), we consider two platforms i and j in a duopoly as

well as two groups of agents, sellers and buyers, denoted respectively s and b.

Agents are uniformly located on a unit segment while platforms are located at each of its extremities. Agents incur a unit transport cost, denoted t. This cost is supposed to be the same for sellers and buyers, which is consistent with our idea that electronic TSM are particularly flexible: whether an agent is seller or buyer does not affect the cost he pays to connect to platforms. For platforms, the total cost of providing the matching service to sellers and to buyers is f_s and f_b respectively.

The agents' utility increases with the number of agents from the other group: a seller's valuation for the participation of one buyer is given by α_s while a buyer's valuation for the participation of one seller is α_b.

Let $p_{s,i}$ (resp. $p_{s,j}$) be the price charged to sellers and $p_{b,i}$ (resp. $p_{b,j}$) the price charged to buyers by the platform i (resp. j). These prices, also called participation fees, are fixed and independent of the outcome of the matching and of the amount of the potential transaction[2]. On eBay, for instance, agents are charged for each transaction they make. But they also pay upstream, when they put an advertisement on-line. It is that kind of fee we consider here.

We now introduce the crucial assumption of our model. It aims at accounting for the specificity of virtual platforms, in which sellers can easily become buyers and *vice versa*. For example, an agent who initially bought a camera on eBay may then buy a CD player or sell a piece of furniture on the same platform. This feature characterizes electronic second hand markets, in which private individuals and collectors participate without any definitive seller's or buyer's statute. In our model the period during which sellers offer their items on-line while buyers browse advertisements is called a matching session, whether a transaction finally takes place or not. We extend Armstrong (2006)'s model by assuming that agents choose one of the two platforms to participate in *two* successive and independent sessions[3]. These two matching sessions are denoted M_1 and M_2 respectively. In the remainder of the paper, we will call initial sellers (resp. initial buyers) agents who are sellers (resp. buyers) during M_1, whatever their type during M_2. We assume that between M_1 and M_2, a proportion β of initial sellers become buyers and that a proportion λ of initial buyers become sellers. These mobility rates are exogeneous.

We finally introduce the following notations. We denote $n^1_{s,i}$ (resp. $n^1_{s,j}$) the

[2]For a model in which electronic platforms play the role of experts about the value of the sellers' good and can manipulate this information to extract profit, see Gaudel & Jullien (2007).

[3]In our model there is no "chicken and egg" problem *à la* Caillaud & Jullien (2003) ("sellers accept to participate if there is a sufficient number and buyers and *vice versa*) since we assume that agents always join a platform.

number of sellers and $n^1_{b,i}$ (resp. $n^1_{b,j}$) the number of buyers on the platform i (resp. j) during M_1. Let $n^2_{s,i}$ (resp. $n^2_{s,j}$) be the number of sellers and $n^2_{b,i}$ (resp. $n^2_{b,j}$) the number of buyers on the platform i (resp. j) during M_2. We assume that $n^1_{s,i} + n^1_{s,j} = n^1_{b,i} + n^1_{b,j} = 1$.

Turning to agents' utilities, we call $U^1_{s,i}$ (resp. $U^1_{s,j}$) the utility obtained by a seller and $U^1_{b,i}$ (resp. $U^1_{b,j}$) the utility obtained by a buyer on the platform i (resp. j) during M_1. We refer to $U^2_{s,i}$ (resp. $U^2_{s,j}$) as the expected utility get by an initial seller on the platform i (resp. j) during M_2 and to $U^2_{b,i}$ (resp. $U^2_{b,j}$) as the expected utility get by an initial buyer on the platform i (resp. j) during M_2. As in Armstrong (2006), we consider that utility functions are linear.

2.2 Timing of actions

The timing of actions is as follows:

1. each agent knows if he will participate in M_1 as a seller or as a buyer. No agent knows if he will remain in his initial group or if he will switch between M_1 and M_2. This assumption illustrates current situations in which one discovers eBay from the buyer or the seller side without wondering if he will remain in the same group or not for the next session. We suppose that all agents and both platforms know the mobility rate of each group.

2. Platforms then set the prices that sellers and buyers will pay respectively for their participation. These fees are set once and for all and do not depend on whether the session is M_1 or M_2. This assumption seems quite realistic: electronic marketplaces set prices for a given period and do not continuously adjust them according to the arrival of new sellers and buyers.

3. Agents then choose once and for all the platform through which they will participate in *both* sessions. This assumption contributes to the originality of our paper since the case in which agents choose a platform *before each session* can be easily solved assuming that agents are involved in a Armstrong (2006)'s game two consecutive times.

4. Sellers and buyers pay their participation fee to their respective platform. Session M_1 takes place.

5. A proportion β of initial sellers become buyers and a proportion λ of initial buyers become sellers.

6. Prices are paid. Session M_2 takes place.

We consider a two-stage game. In the first stage, the two platforms compete in prices. In the second stage, each agent chooses its platform for the two sessions.

3 Equilibrium

3.1 The second-stage subgame: agents choose between the two platforms

We first examine the agents' choice between the two platforms. During M_1, agents' utilities are as follows:

$$
\begin{aligned}
.U^1_{s,i} &= (\alpha_s n^1_{b,i} - p_{s,i}) \\
U^1_{s,j} &= (\alpha_s n^1_{b,j} - p_{s,j}) \\
U^1_{b,i} &= (\alpha_b n^1_{s,i} - p_{b,i}) \\
U^1_{b,j} &= (\alpha_b n^1_{s,j} - p_{b,j})
\end{aligned}
$$

The agents' expected utility during M_2 depends on whether they switch or not. For example, if an initial seller does not switch, he enjoys buyers' participation in M_2; if he switches, he values sellers' participation. We thus have :

$$
\begin{aligned}
U^2_{s,i} &= (1-\beta)(\alpha_s n^2_{b,i} - p_{s,i}) + \beta(\alpha_b n^2_{s,i} - p_{b,i}) \\
U^2_{s,j} &= (1-\beta)(\alpha_s n^2_{b,j} - p_{s,j}) + \beta(\alpha_b n^2_{s,j} - p_{b,j}) \\
U^2_{b,i} &= (1-\lambda)(\alpha_b n^2_{s,i} - p_{b,i}) + \lambda(\alpha_s n^2_{b,i} - p_{s,i}) \\
U^2_{b,j} &= (1-\lambda)(\alpha_b n^2_{s,j} - p_{b,j}) + \lambda(\alpha_s n^2_{b,j} - p_{s,j})
\end{aligned}
$$

Since agents are supposed to choose their platform for the two sessions, their choice depends on total expected utility, denoted and defined as follows:

$$
\begin{aligned}
E(U_{s,i}) &\equiv U^1_{s,i} + E(U^2_{s,i}) \\
E(U_{s,j}) &\equiv U^1_{s,j} + E(U^2_{s,j}) \\
E(U_{b,i}) &\equiv U^1_{b,i} + E(U^2_{b,i}) \\
E(U_{b,j}) &\equiv U^1_{b,j} + E(U^2_{b,j})
\end{aligned}
$$

It follows that

$$
\begin{aligned}
E(U_{s,i}) &= (\alpha_s n^1_{b,i} - p_{s,i}) + (1-\beta)(\alpha_s n^2_{b,i} - p_{s,i}) + \beta(\alpha_b n^2_{s,i} - p_{b,i}) && (1) \\
E(U_{s,j}) &= (\alpha_s n^1_{b,j} - p_{s,j}) + (1-\beta)(\alpha_s n^2_{b,j} - p_{s,j}) + \beta(\alpha_b n^2_{s,j} - p_{b,j}) && (2) \\
E(U_{b,i}) &= (\alpha_b n^1_{s,i} - p_{b,i}) + (1-\lambda)(\alpha_b n^2_{s,i} - p_{b,i}) + \lambda(\alpha_s n^2_{b,i} - p_{s,i}) && (3) \\
E(U_{b,j}) &= (\alpha_b n^1_{s,j} - p_{b,j}) + (1-\lambda)(\alpha_b n^2_{s,j} - p_{b,j}) + \lambda(\alpha_s n^2_{b,j} - p_{s,j}) && (4)
\end{aligned}
$$

Since they choose their platform before M_1 once and for all, agents incur the transport cost according to their initial type, whatever their type during M_2[4]. As in Armstrong (2006), we use the well known Hotelling specification to determine agents' participation. We obtain

$$n^1_{s,i} = \frac{1}{2} + \frac{E(U_{s,i}) - E(U_{s,j})}{2t} \quad (5)$$

$$n^1_{s,j} = \frac{1}{2} + \frac{E(U_{s,j}) - E(U_{s,i})}{2t} \quad (6)$$

$$n^1_{b,i} = \frac{1}{2} + \frac{E(U_{b,i}) - E(U_{b,j})}{2t} \quad (7)$$

$$n^1_{b,j} = \frac{1}{2} + \frac{E(U_{b,j}) - E(U_{b,i})}{2t} \quad (8)$$

We now determine the number of sellers and of buyers on each platform during M_2. Contrary to a situation with no mobility, the population of sellers and of buyers evolves between M_1 and M_2. For example, the population of sellers during M_2 results from the switching of some initial buyers and from the immobility of some initial sellers. We thus have

$$n^2_{s,i} = (1-\beta)n^1_{s,i} + \lambda n^1_{b,i} \quad (9)$$

$$n^2_{s,j} = (1-\beta)n^1_{s,j} + \lambda n^1_{b,j} \quad (10)$$

$$n^2_{b,i} = (1-\lambda)n^1_{b,i} + \beta n^1_{s,i} \quad (11)$$

$$n^2_{b,j} = (1-\lambda)n^1_{b,j} + \beta n^1_{s,j} \quad (12)$$

3.2 The first-stage subgame: platforms compete in prices

Platforms set equilibrium prices $p_{s,i}*$, $p_{b,i}*$, $p_{s,j}*$ and $p_{b,j}*$ according to the following program:

$$\begin{cases} \{p^*_{s,i}, p^*_{b,i}\} = \text{ArgMax}\Pi_i = \text{ArgMax}(p_{s,i} - f_s)(n^1_{s,i} + n^2_{s,i}) + (p_{b,i} - f_b)(n^1_{b,i} + n^2_{b,i}) \\ \{p^*_{s,j}, p^*_{b,j}\} = \text{ArgMax}\Pi_j = \text{ArgMax}(p_{s,j} - f_s)(n^1_{s,j} + n^2_{s,j}) + (p_{b,j} - f_b)(n^1_{b,j} + n^2_{b,j}) \end{cases} \quad (13)$$

[4]Alternatively, we can assume that transport costs are not paid when agents choose their platform (i.e. once and for all) but each time they connect and participate in a session (i .e. before M_1 *and* before M_2). This requires to determine, for each initial type of agents, an expected transport cost over the two sessions depending on mobility rates. But in so far as transport costs are the same for sellers and for buyers, this does not qualitatively affect our results.

where Π_i and Π_j denote plateforms' expected profits over the two sessions.

We obtain the following proposition:

Proposition 1. *For a certain range of parameters, there exists an equilibrium. It is unique and symmetric. Equilibrium prices are given by*

$$p^*_{s,i} = p^*_{s,j} = p^*_s = \hat{p}^*_s + \Delta p^*_s$$

$$p^*_{b,i} = p^*_{b,j} = p^*_b = \hat{p}^*_b + \Delta p^*_b$$

$$\text{with } \hat{p}^*_s = f_s + \frac{1}{2}t - \alpha_b,\ \hat{p}^*_b = f_b + \frac{1}{2}t - \alpha_s,$$

$$\Delta p^*_s = \frac{-\beta\frac{1}{2}(\lambda-\beta)\alpha_s - 2(1-\beta)\frac{1}{2}(\beta-\lambda)\alpha_b + \lambda\frac{1}{2}(\beta-\lambda)\alpha_b}{(2-\beta-\lambda)},$$

$$\text{and } \Delta p^*_b = \frac{-\lambda\frac{1}{2}(\beta-\lambda)\alpha_b - 2(1-\lambda)\frac{1}{2}(\lambda-\beta)\alpha_s + \beta\frac{1}{2}(\lambda-\beta)\alpha_s}{(2-\beta-\lambda)}$$

Proof. See Appendix.

Since our equilibrium is symmetric, we restrict our discussion to p^*_s. To see the intuition behind Proposition 1, it is important to note that $\hat{p}^*_s$ is the equilibrium price without mobility ($\lambda = \beta = 0$)[5] while Δp^*_s represents the effect of the mobility assumption on the equilibrium[6]. As explained by Armstrong (2006), $\hat{p}^*_s$ depends on the buyers' externality coefficient: the more the participation of sellers is valued by buyers, the lower the fee charged to sellers. The comment of Δp^*_s is slighly more complex. We will consider successively its numerator and its denominator.

Let us first consider the numerator of Δp^*_s. It is noteworthy that the term $\frac{1}{2}(\lambda-\beta)$ is the (positive or negative) number of sellers on each platform during M_2, *relatively to* a situation without mobility. It measures the size of what we will call the *differential* population of sellers. Symmetrically, $\frac{1}{2}(\beta-\lambda)$ is the (positive or negative) number of buyers on each platform during M_2, *relatively to* a situation

[5] p^*_s is not exactly Armstrong (2006)'s equilibrium price since we consider 2 sessions. Under the assumption that transport costs are paid twice (before M_1 and before M_2), we obtain Armstrong (2006)'s equilibrium prices: $\hat{p}^*_s = f_s + t - \alpha_b$, $\hat{p}^*_b = f_b + t - \alpha_s$.

[6] We find the same kind of decomposition in the case of a monopoly platform.

without mobility. It measures the size of the *differential* population of buyers[7]. So when $\lambda - \beta > 0$, the number of sellers during M_2 is higher than without mobility while the number of buyers is lower. When $\lambda - \beta < 0$, the number of sellers during M_2 is lower than without mobility while the number of buyers is higher. For the clarity purposes, in the remaining of our comment, we will focus on the case where $\lambda - \beta > 0$ (the symmetric reasoning applies if $\lambda - \beta < 0$).

Each component of the numerator of Δp_s^* can be interpreted in terms of reward or of penalty, relatively to the pricing without mobility. A negative sign means that the component tends to reduce the equilibrium price, which represents a reward for agents. On the contrary, a positive sign means that the term tends to increase the equilibrium price, which represents a penalty. The two first components can be understood by considering the case of an initial seller:

- if he becomes buyer (with probability β), his participation will be enjoyed during M_2 by sellers. So, the term $\frac{1}{2}\beta(\lambda-\beta)\alpha_s$ refers to the expected utility get by the differential population of sellers from the participation and the possible switching of initial sellers. If $\lambda - \beta > 0$, then $-\frac{1}{2}\beta(\lambda - \beta)\alpha_s < 0$. This means that sellers are rewarded, which is consistent with the fact that they will exert externalities on a population (the sellers) which is larger than without mobility.

- if the initial seller remains seller (with probability $1 - \beta$), his participation is enjoyed by buyers. Hence, the term $2\frac{1}{2}(1 - \beta)(\beta - \lambda)\alpha_b$ is the expected utility get by the differential population of buyers from the participation and the possible immobility of initial sellers. If $\lambda - \beta > 0$, then $-2\frac{1}{2}(1 - \beta)(\beta - \lambda)\alpha_b > 0$. This means that sellers are penalized: they will exert externalities on a group which is smaller than without mobility.

The third component is less intuitive since it does not bear on the externalities exerted by sellers. The term $\frac{1}{2}\lambda(\beta - \lambda)\alpha_b$ refers to the expected utility obtained by the differential population of buyers from the participation of buyers and from their possible switching into sellers. To understand the meaning of this component, it is important to note that it appears with the opposite sign in Δp_b^*, the price charged to buyers. It can thus be interpreted as a transfer between sellers and buyers. In the expression of Δp_b^*, the term $-\frac{1}{2}\lambda(\beta-\lambda)\alpha_b$ is positive: buyers are penalized since

[7]On the one hand, the proof of Proposition 1 in Appendix indicates that the equilibrium number of initial sellers and of initial buyers on each platform equals $\frac{1}{2}$. Hence, during M_2, the equilibrium number of sellers on each platform is $\frac{1}{2}(1 - \beta + \lambda)$ while the equilibrium number of buyers is $\frac{1}{2}(1-\lambda+\beta)$. On the other hand, without mobility, the number of sellers and of buyers on each platform during M_2 is the same as during M_1 i.e. equal to $\frac{1}{2}$. Making a simple substraction is enough to show that $\frac{1}{2}(\lambda - \beta)$ (resp. $\frac{1}{2}(\beta - \lambda)$) is the (positive or negative) equilibrium number of sellers (resp. buyers) on each platform during M_2, *relatively to* a situation without mobility.

they will exert externalities on a population (the buyers) which is smaller than without mobility. In the expression of Δp_s^*, $\frac{1}{2}\lambda(\beta-\lambda)\alpha_b 0$ is negative, which means that thanks to the buyers' penalty, platforms can afford to charge a lower price to sellers.

We now turn to the denominator of Δp_s^*. While $\beta+\lambda$ measures global mobility, $2-\beta-\lambda$ indicates the proportion of the population which is steady between the two sessions. The denominator can be seen as a factor that adjusts the strenght of the three mechanisms described above: when global mobility increases their intensity is amplified.

Finally, our analysis of Δp_s^* points out two interesting phenomena. On the one hand, our findings indicate that mobility does not have a trivial effect on equilibrium. As rewards do not exactly compensate penalties, equilibrium prices can be higher or lower than prices without mobility. On the other hand, owing to their own or to other's mobility, agents are led to value the participation of agents belonging to their initial group. Before M_1, agents do not know if they will migrate or not to the other group. Since they may switch, they enjoy the participation of the moving part of their initial group ; since they may also not switch, they enjoy the participation of its steady part too.

From equilibrium prices it is straightforward to derive the following proposition:

Proposition 2 *Platforms' equilibrium profit is given by*

$$\Pi_i^* = \Pi_j^* = \Pi^* = 2\hat{\Pi}^* + 2\Delta\Pi^*$$

$$\textit{with } \hat{\Pi}^* = \frac{t-\alpha_s-\alpha_b}{2}, \textit{ and } \Delta\Pi^* = \frac{-(1-\lambda)\frac{1}{2}(\lambda-\beta)\alpha_s - (1-\beta)\frac{1}{2}(\beta-\lambda)\alpha_b}{(2-\beta-\lambda)}$$

Proposition 2 states that the platforms' equilibrium profit can be written as the sum of two terms. $\hat{\Pi}^*$ is the equilibrium profit without mobility[8]. As noticed by Armstrong (2006), it decreases with agents' externality coefficients.

[8]As in Proposition 1, $\hat{\Pi}^*$ is not exactly Armstrong (2006)'s equilibrium profit. Under the assumption that transport costs are paid before M_1 and before M_2, we find Armstrong (2006)'s equilibrium profit:

$$\hat{\Pi}^* = \frac{2t-\alpha_s-\alpha_b}{2}.$$

The term $\Delta\Pi^*$ represents the effect of the mobility assumption on the equilibrium profit. Its denominator can be interpreted in the same way as in Proposition 1. The numerator of $\Delta\Pi^*$ has two components, both depending on the agents' probability to keep their initial type: $(1-\lambda)$ for sellers and $(1-\beta)$ for buyers. The term $-(1-\lambda)\frac{1}{2}(\lambda-\beta)\alpha_s$ refers to the expected utility get by the differential population of sellers from the participation and the possible immobility of buyers. The term $-(1-\beta)\frac{1}{2}(\beta-\lambda)\alpha_b$ refers to the expected utility get by the differential population of buyers from the participation and the possible immobility of sellers. If $\lambda-\beta>0$, buyers are rewarded since they will exert externalities on a population (the sellers) which will be larger than without mobility. This represents a loss for platforms, which is consistent with the fact that $-(1-\beta)\frac{1}{2}(\lambda-\beta)\alpha_s<0$. Symmetrically, sellers are penalized since they will exert externalities on a population (the buyers) which will be smaller than without mobility. This constitutes a profit for platforms. This is consistent with the fact that $-(1-\lambda)\frac{1}{2}(\beta-\lambda)\alpha_b>0$. The discussion is reversed if $\lambda-\beta<0$.

Whether Π^* is higher or lower than the equilibrium profit without mobility depends on the sign of $\Delta\Pi^*$. It is straightforward that $\Delta\Pi^*>0$ if $(\lambda+\beta)>\xi$ with $\xi\equiv\frac{(\lambda-\beta)(\alpha_s+\alpha_b)}{\alpha_b-\alpha_s}+2$. This implies that for sufficiently high (resp. low) values of global mobility, the equilibrium profit is higher (resp. lower) than without mobility. Note that, since $\lambda+\beta$ is always higher than 2, the condition for having $\Delta\Pi^*>0$ is never satistied when $\xi>2$ i.e. when $\frac{(\lambda-\beta)(\alpha_s+\alpha_b)}{\alpha_b-\alpha_s}>0$. This means that Π^* is always lower than the equilibrium profit without mobility when the group having the highest externality parameter also has the lowest mobility rate ($\lambda-\beta>0$ and $\alpha_s-\alpha_b>0$ or $\lambda-\beta>0$ and $\alpha_s-\alpha_b>0$).

Finally, our result shows that platforms earn profit on the population that is larger than without mobility and make losses on the population that is smaller. It is noteworthy that these effects are not symmetric, such that the game "reward/penalty" is not a zero-sum game.

3.3 Comparative statics analysis

Let us now turn to the comparative statics of the equilibrium. We first investigate the impact of a change in agents' individual mobility. As there exist two types of agents, it seems more interesting to focus on relative mobility (the mobility of a group relatively to the one of the other group) instead of the absolute mobility of a given group.

Proposition 3.
(a) There exists a threshold μ^* such that

If $-1 < \lambda - \beta < \mu^*$, equilibrium prices decrease with $\lambda - \beta$,

If $\mu^* < \lambda - \beta < 1$, equilibrium prices increase with $\lambda - \beta$.

(b) *There exists a threshold μ^{**} such that*

If $-1 < \lambda - \beta < \mu^{**}$, the platforms' profit decreases with $\lambda - \beta$,

If $\mu^{**} < \lambda - \beta < 1$, the platforms' profit increases with $\lambda - \beta$.

Proof. See Appendix.

Proposition 3 establishes that prices and the platforms' profit depend on relative mobility. It suggests that platforms capture agents' heterogeneity in terms of mobility to set high prices and to extract profit.

According to Part (a) of Proposition 3, the minimum of equilibrium prices is reached for $\lambda - \beta = \mu^*$. When $\lambda - \beta$ moves away from this value, equilibrium prices rise. They reach their maximum when $\lambda - \beta$ tends to -1 or to 1. This is illustrated in Graph 1, in which the set of parameters is $\lambda + \beta = 1$, $\alpha_s = 2$ and $\alpha_b = 3$[9].

[9]Note that, using bound expressions given in the proof of Proposition 3, we obtain the following property: if $\lambda + \beta < \frac{4\alpha_b}{\alpha_s + 3\alpha_b}$, the maximum of equilibrium profit is reached when $\lambda - \beta$ tends to 1 and if $\lambda + \beta > \frac{4\alpha_b}{\alpha_s + 3\alpha_b}$, it is reached when $\lambda - \beta$ tends to -1.

The intuition behind Part (a) is as follows. Agents, sellers for instance, are charged the highest price if, relatively to the equilibrium without mobility, they are much more penalized than rewarded by platforms. The penalty effect is predominant if sellers have a *high* probability to exert externalities during M_2 on the population which will be smaller than without mobility and a *low* probability to exert externalities on the population which will be larger than without mobility. It is precisely what happens when the difference in mobility rates is high, i.e. when λ tends to 1 while β tends to 0 or when λ tends to 0 while β tends to 1.

To see that in detail, let us focus on the former case. When λ tends to 1 while β tends to 0, initial sellers are very likely to remain sellers. Their probability to exert externalities during M_2 on buyers (who will be less numerous than without mobility since $\lambda - \beta > 0$), is very high while their probability to exert externalities on sellers (who will be more numerous than without mobility) is very low. The former effect thus dominates the latter. In this case, the main determinant of Δp_s^* is a penalty, measured by the expected utility get by the differential population of buyers from the participation and the immobility of initial sellers. Hence, the second term of the numerator of Δp_s^* should be positive and larger in absolute value than the sum of the two other terms. This is well illustrated in the following numerical example. Setting $\lambda = 0.99$, $\beta = 0.01$, $\alpha_s = 3$ and $\alpha_b = 2$, we obtain: $-\frac{1}{2}\beta(\lambda-\beta)\alpha_s = -0.014$, $-2\frac{1}{2}(1-\beta)(\beta-\lambda)\alpha_b = 1.940$ and $\frac{1}{2}\lambda(\beta-\lambda)\alpha_b = -0.970$.

When λ tends to 0 and β tends to 1, the reasonning is symmetric. The main determinant of Δp_s^* is the expected utility get by the differential population of buyers from the participation and the possible switching of initial sellers. Using the same set of parameters as above but reversing the values of λ and β (such that the denominator is unchanged), we confirm that the first term of the numerator is positive and larger in absolute value than the sum of the two other terms: $-\frac{1}{2}\beta(\lambda-\beta)\alpha_s = 1.455$, $-2\frac{1}{2}(1-\beta)(\beta-\lambda)\alpha_b = -0.019$ and $\frac{1}{2}\lambda(\beta-\lambda)\alpha_b = 0.098$.

Part (b) of Proposition 3 states that the profit maximum is associated with the largest difference in mobility rates. This is depicted in the plot below, in which the set of parameters is: $\lambda + \beta = 1.3$, $\alpha_s = 2$ and $\alpha_b = 4$.

To understand the intuition of part (b), remind that the platforms' equilibrium profit depends on two terms: the expected utility get by the differential population of sellers from the possible immobility of buyers and the expected utility get by buyers from the possible immobility of sellers. What happens when the difference in mobility rates is very high? As above, let us focus on the case where λ tends to 1 while β tends to 0. On the one hand, since $\lambda - \beta > 0$, buyers are rewarded. This implies a loss for platforms. But the expected value of this loss is low since initial buyers are very unlikeky to remain buyers and to exert externalities on sellers (λ tends to 1). On the other hand, since $\beta - \lambda < 0$, sellers are penalized. This constitutes a profit for platforms. The expected value of this profit is high since initial buyers are very likeky to remain sellers and to exert externalities on buyers (β tends to 0). Finally, the latter effect dominates the former such that the platforms' profit is high. The first term of the denominator of $\Delta\Pi^*$ should be negative while the second term should be positive and larger in absolute value. For example, when $\lambda = 0.99$, $\beta = 0.01$, $\alpha_s = 3$ and $\alpha_b = 2$, we have: $-(1-\lambda)(\lambda-\beta)\alpha_s = -0.294$ and $-(1-\beta)(\beta-\lambda)\alpha_b = 1.940$. When λ tends to 0 while β tends to 1, the result is symmetric. For example, when $\lambda = 0.01$, $\beta = 0.99$, we obtain: $-(1-\lambda)(\lambda-\beta)\alpha_s = 2.910$ and $-(1-\beta)(\beta-\lambda)\alpha_b = -0.190$.

Studying the impact of a shift in global mobility, we obtain the following proposition:

Proposition 4. The platforms' equilibrium profit increases with $\lambda + \beta$.

Proof. See Appendix.

Proposition 4 reports that both platforms make additional profit when global mobility rises. This phenomenon actually results from two effects. On the one hand, sellers and buyers are less likely to keep their initial type during M_2. Hence, the reward as well as the penalty effect captured by the numerator of $\Delta\Pi^*$ tend to vanish since, as mentionned in the discussion of Proposition 2, they both stem from the possible immobility of agents. But depending on parameter values, the fall in the reward may dominate the fall in the penalty or *vice-versa* such that the effect of global mobility on the numerator is undetermined. For example, setting $\alpha_s = 3$ and $\alpha_b = 2$, the numerator equals 0.055 if $\lambda = 0.1$ and $\beta = 0.2$ and decreases to 0.020 when $\lambda = 0.8$ and $\beta = 0.9$. Inversely, setting $\alpha_s = 2$, $\alpha_b = 3$, the numerator equals -0.03 when $\lambda = 0.2$ and $\beta = 0.2$ and increases to 0.005 when $\beta = 0.8$ and $\lambda = 0.9$.

On the other hand, as global mobility increases, the denominator of $\Delta\Pi^*$ decreases. As explained in our comment of Proposition 1, a larger global mobility implies a stronger amplification of the effects captured by the numerator.

Our result suggests that firms can conduct industrial strategies in order to promote switching from one group to another. In a transaction fee regime, a change in global mobility stimulates business volume and profit earning. This trivial effect probably prevails in a majority of TSM with group-switching. For example, eBay regularly urges buyers to become sellers for the next transaction. Proposition 4 reveals a more subtle effect that appears even in a participation fee regime with a fixed number of transactions (i.e. for the same volume of business). The simple fact that a buyer may become a seller (or *vice-versa*) gives the platform an opportunity to make profit.

Considered together, Proposition 3 and Proposition 4 indicate that the overall impact of mobility is based on a complex overlapping between global and relative mobility. This suggests being very carefull when analysing the effects of a shift in mobility. The decrease in the mobility of one of the two groups reduces global mobility but it also has an ambiguous effect on relative mobility such that the overall effect on profit is not trivial.

4 Conclusion

In this paper, we define Internet platforms as two-sided markets in which both sides of the market can easily switch. To account for this specificity, we consider two platforms in a duopoly through which sellers and buyers match during two successive sessions. The main contribution of our paper is to assume that each group is characterized by an exogeneous mobility rate such that between the two sessions, some sellers become buyers and *vice-versa*.

We show that plateforms' equilibrium prices can be written as the sum of two terms. The first one is the equilibrium price without mobility. The second one can be interpreted in terms of rewards/penalties relatively to the equilibrium without mobility. As rewards do not perfectly compensate penalties, equilibrium prices can be higher or lower than prices without mobility. We also identify global and relative mobility as sources of platforms' profit.

Electronic intermediation is a fruitful topic which deserves further investigation. Our analysis could be refined by assigning an endogeneous component to agents' mobility. Another interesting (and possibly complementary) development could consist in accounting for competition among agents belonging to the same side of the market, as in Belleflamme & Toulemonde (2007).

Appendix

Proof of Proposition 1. Owing to the complexity of the maximization program, we use the computation software Maple to solve the model analytically. To avoid useless heaviness, we present the main steps of the maximisation process without expliciting their analytical expression. Inserting Eqs. (1), (2), (3) and (4) into Eqs. (5), (6), (7) and (8) yields Eqs. (5'), (6'), (7') and (8'). Inserting them into Eqs. (9), (10), (11) and (12), we obtain (9'), (10'), (11') and (12'). Eqs. (5'), (6'), (7') and (8') and Eqs. (9), (10), (11) and (12) are finally inserted in (13) to obtain (13'). We then derive (13') with respect to prices and we compute first-order conditions. For β et λ simultaneously different from 1, we obtain:

$$p_{s,i}^* = p_{s,j}^* = p_s^* = \hat{p}_s^* + \Delta p_s^*$$

$$p_{b,i}^* = p_{b,j}^* = p_b^* = \hat{p}_b^* + \Delta p_b^*$$

$$\text{with } \hat{p}_s^* = f_s + \frac{1}{2}t - \alpha_b,\ \hat{p}_b^* = f_b + \frac{1}{2}t - \alpha_s,$$

$$\Delta p_s^* = \frac{1}{2}(\lambda - \beta)\frac{-\beta\alpha_s - 2(\beta - \lambda)\alpha_b + \lambda\alpha_b}{(2 - \beta - \lambda)},$$

$$\Delta p_b^* = \frac{1}{2}(\lambda - \beta)\frac{-\lambda\alpha_b - 2(1 - \lambda)\alpha_s + \beta\alpha_s}{(2 - \beta - \lambda)}$$

Equilibrium agents' participation is given by:

$$n_{s,i}^{1*} = n_{s,j}^{1*} = n_{b,i}^{1*} = n_{b,j}^{1*} = \frac{1}{2}$$

We also ensure that the second-order conditions are satisfied. We find that there exists a certain range of parameters for which the determinant of the Hessian matrix is positive[10].

Proof of Proposition 3.

Part (a). Owing to the symmetry of the equilibrium, we limit our attention to p_s^*. Since $\hat{p}_s^*$ does not depend on β and λ, we focus on Δp_s^*. We use an invertible change of variable to study this term. Setting $\theta = \lambda + \beta$ and $\mu = \lambda - \beta$ with $0 \leq \theta < 2$ and $-1 < \mu < 1$, we have $\beta = \frac{\theta+\mu}{2}$ and $\lambda = \frac{\theta-\mu}{2}$. It follows that

$$\Delta p_s^* = \Psi(\theta, \mu) = \frac{\mu(\alpha_s\theta - \alpha_s\mu + 3\alpha_b\theta - \alpha_b\mu - 4\alpha_b)}{4(\theta - 2)}.$$

Deriving $\Psi(\theta, \mu)$ with respect to μ yields

$$\frac{\delta\Psi(\theta, \mu)}{\delta\mu} = \frac{\alpha_s\theta - 2\alpha_s.\mu + 3\alpha_b\theta - 2\alpha_b\mu - 4\alpha_b}{4(\theta - 2)}.$$

It is straighforward to show that

$$\frac{\delta\Psi(\theta, \mu)}{\delta\mu} = 0 \text{ for } \mu = \mu^* \equiv \frac{\alpha_s\theta + 3\alpha_b\theta - 4\alpha_b}{\alpha_s + \alpha_b}.$$

Therefore if $-1 < \mu < \mu^*$, then $\frac{\delta\Psi(\theta, \mu)}{\delta\mu} < 0$ and if $\mu^* < \mu < 1$, then $\frac{\delta\Psi(\theta, \mu)}{\delta\mu} > 0$

$$\text{with } \lim_{\mu \to -1} \Delta p_s^* = \lim_{\mu \to 1} \Delta p_b^* = \frac{-3\alpha_b\theta + 3\alpha_2 - \alpha_1 - \alpha_s\theta}{4(\theta - 2)}$$

[10]The determinant of the Hessian matrix is positive for sufficiently large values of t relatively to α_s and α_b.

$$\text{and } \lim_{\mu\to 1}\Delta p_s^* = \lim_{\mu\to -1}\Delta p_b^* = \frac{3\alpha_b\theta - 5\alpha_2 - \alpha_1 + \alpha_s\theta}{4(\theta-2)}.$$

Part (b). Resorting to the same invertible change of variable as above, we obtain

$$\Delta\Pi^* = \Phi(\theta,\mu) = \frac{\mu(-2\alpha_b - \alpha_b\mu - \alpha_s\theta + 2\alpha_s - \alpha_s\mu + \alpha_b\theta)}{2(\theta-2)}.$$

It follows that

$$\frac{\delta\Phi(\theta,\mu)}{\delta\mu} = \frac{-2\alpha_b - 2\alpha_b\theta - 2\alpha_s\mu + 2\alpha_s - \alpha_s\theta + \alpha_b\theta}{2(\theta-2)}$$

$$\text{and } \frac{\delta\Phi(\theta,\mu)}{\delta\mu} = 0 \text{ for } \mu = \mu^{**} \equiv \frac{\mu(\theta-2)}{2\theta}.$$

Hence if $-1 < \mu < \mu^{**}$, then $\frac{\delta\Phi(\theta,\mu)}{\delta\mu} < 0$ and if $\mu^{**} < \mu < 1$ then $\frac{\delta\Phi(\theta,\mu)}{\delta\mu} > 0$

$$\text{with } \lim_{\mu\to -1}\Delta\Pi^* = \frac{\alpha_b - 3\alpha_s + \alpha_s\theta - \alpha_b\theta}{2(\theta-2)} \text{ and } \lim_{\mu\to 1}\Pi^* = \frac{3\alpha_b - \alpha_s + \alpha_s\theta - \alpha_b\theta}{2\theta - 2}.$$

Proof of Proposition 4. We start from $\Phi(\theta,\mu)$ defined in the proof of Proposition 3. Deriving $\Phi(\theta,\mu)$ with respect to θ yields

$$\frac{\delta\Phi(\theta,\mu)}{\delta\theta} = \frac{\mu^2(\alpha_s+\alpha_b)}{2(\theta-2)^2}$$

.

It is straighforward that

$$\frac{\delta\Phi(\theta,\mu)}{\delta\theta} > 0.$$

References

Anderson, S. & Gabszewicz, J. (2006), *The media and advertising: a tale of two-sided markets in Handbook of the Economics of Art and Culture*, V. Ginsburgh and D. Throsby (eds.), Elsevier.

Armstrong, M. (2006), 'Competition in two-sided markets', *Rand Journal of Economics* **37**, 668–691.

Belleflamme, P. & Toulemonde, E. (2007), 'Negative intra-group externalities in two-sided markets', *CESifo Working paper 2011*.

Caillaud, B. & Jullien, B. (2001), 'Competing cybermediaries', *The European Economic Review* **45**, 797–808.

Caillaud, B. & Jullien, B. (2003), 'Chicken and egg: Competition among intermediaries services providers', *The Rand Journal of Economics* **34**, 309–328.

Chakravorti, S. & Roson, R. (2006), 'Platforms competition in two-sided markets: The case of payment networks', *Review of Network Economics* **5**, 118–142r.

Dinlersoz, E. & Pereira, P. (2007), 'On the diffusion of electronic commerce', *International Journal of Industrial Organization* **25**, 541–574.

Evans, D. (2003), 'Some empirical aspects of multi-sided platform industries', *Review of Network Economics* **2**, 191–209.

Ferrando, J., Gabszewicz, J., Laussel, D. & Sonnac, N. (2004), 'Two-sided network effect and competition: An application to media industries', *mimeo*.

Gabszewicz, J., Laussel, D. & Sonnac, N. (2001), 'Press advertising and the ascent of the pensée unique', *European Economic Review* **45**, 654–651.

Gabszewicz, J., Laussel, D. & Sonnac, N. (2004), 'Press advertising and political differentiation of newspapers', *Journal of Public Economic Theory* **4**, 249–259.

Gaudel, A. & Jullien, B. (2007), E-commerce, Two-Sided Markets and Info-Mediation, *in Internet and Digital Economics*, E. Brousseau and N. Curien, Cambridge University Press.

Guthrie, G. & Wright, J. (2003), 'Competition payment schemes', *Working paper 311, Departments of Economics, National University of Singapore*.

Kaiser, U. & Wright, J. (2006), 'Price structure in two-sided markets: Evidence from the magazine industry', *International Journal of Industrial Organization* **24**, 1–28.

Rochet, J. & Tirole, J. (2002), 'Cooperation among competitors: the economics of payment card association', *Rand Journal of Economics* **33**, 549–570.

Rochet, J. & Tirole, J. (2003), 'Platform competition in two-sided markets', *Journal of European Economic Association* **1**, 990–1029.

Rochet, J.-C. & Tirole, J. (2006), 'Two-sided markets: A progress report', *Rand Journal of Economics* **37**, 645–667.

Roson, R. (2005), 'Two-sided markets: A tentative survey', *Review of Network Economics* **4**, 142–160.

Private Cards and the Bypass of Payment Systems by Merchants

Marc Bourreau*and Marianne Verdier†

June 13, 2008

Abstract

This paper studies the incentives of a merchant to enter the market for payment card transactions by issuing private cards. In our setting, two merchants that are differentiated à la Hotelling compete on the product market. A payment platform organizes the interactions between a monopolistic issuer and a monopolistic acquirer by choosing a level of interchange fee. We show that, if a merchant issues private cards, he sets a very aggressive price to compete with the issuer. The competition with the private card generates a fall in the payment card fee and a rise in the merchant fee. We show that, in our setting, the payment platform may decide to increase the level of the interchange fee in order to deter the merchant from entering the market for payment transactions.

JEL Codes: G21, L31, L42.

Keywords: Payment card systems, interchange fee, two-sided markets, private cards.

*Institut TELECOM, TELECOM ParisTech, Department of Economics and Social Sciences, and CREST-LEI, Paris. Email: marc.bourreau@telecom-paristech.fr

†Institut TELECOM, TELECOM ParisTech, Department of Economics and Social Sciences, Paris. Tel: 33 1 53 89 35 09. Email: marianne.verdier@telecom-paristech.fr

1 Introduction

In the United-States, in 2006, payment card transactions cost merchants nearly $57 billion.[1] The costs of card payments is a major source of conflict between banks and merchants. Merchants have to pay a fee (the "merchant fee") to their bank (the "Acquirer") each time a consumer pays by card, which they claim to be excessive.[2] In 2005, in the United-States, the usual amount of the merchant fee ranged from 1% to 2.7% of the transaction. Merchants argue that they cannot pass through to consumers the cost of a payment card transaction, since surcharges are forbidden by most payment card associations (like Visa). Also, they contend that it has become impossible to refuse a payment instrument which is now widely used by consumers.

This explains why merchants have thought about strategies to reduce the costs of payment card transactions. One of these strategies, which has been implemented by large retailers such as Wal-Mart or JC Penney and Macy's, has been to start issuing "private cards". Unlike payment cards issued by banks, which are members of payment card associations, private cards can only be used at the retailer's shop. The private card enables the merchant to save the cost of the merchant fee, if it is issued without the support of a financial institution. The detention and usage of private cards have become widespread over the last ten years. According to the International Card Manufacturer's Association (IMCA), 5.6 billion private cards have been sold or delivered worldwide in 2004. Private cards account for 42.9% of the cards issued.

The purpose of this paper is to analyse merchants' incentives to issue their private cards, and to characterise the possible reactions of the payment card association.

Payment card networks are often managed by a payment association, such as Visa, or MasterCard, which organises the interactions between the bank of the cardholder, the "Issuer", and the bank of the merchant, the "Acquirer". Such payment card associations entail several benefits for the bank members. For instance, an Issuer is ensured that his payment card will be accepted by all the merchants that are affiliated with the association, while an Acquirer knows that the Issuer will respect the rules for security and processing that are designed by the association. Hence, banks benefit from network effects of membership, and from a reduction in the asymmetry of information when they proceed to payment transactions. Payment card associations also enable banks to allocate optimally the total cost of a payment card transaction between each other, by choosing an "interchange" fee, that is paid by the Acquirer to the Issuer,

[1] Source: Nilson Report, Issue 877 (2007).

[2] See for instance www.nationalgrocers.org, "Of further concern is the grocery industry trend toward both higher interchange rates and higher volumes of electronic transactions, with a number of companies reporting more than 50 percent of their purchases being made with credit and debit cards". See also the Visa Wal-Mart case (2003).

each time a consumer pays by card. The effect of the interchange fee is to reduce the marginal cost of the Issuer and to increase the marginal cost of the Acquirer. This is a way for the payment association of subsidizing the consumers' side, by allowing the Issuers to choose a lower price for the payment card, to the detriment of the merchants' side. Hence, though interchange fees stimulate the demand for card payments, their effect on merchants' side may provide large retailers with incentives to bypass the payment card association.

This paper aims also at analysing the impact of private cards on the level of the interchange fee that is chosen by the payment association. We try to determine if the payment platform can use the interchange fee to deter merchants from entering the market for payment card transactions. As we will show, this issue is not trivial.

The possibility to bypass the payment association by issuing private cards has never been studied in the literature on payment card systems.[3] Among others, Rochet and Tirole (2002) and Wright (2004) show that the optimal level of interchange fee depends on the nature of competition between merchants. An interesting insight, provided by Rochet and Tirole (2002), is that merchants are ready to accept higher merchant fees to avoid losing market share if they refuse cards. But no paper takes into account the fact that merchants can compete with the payment association by providing their own payment services.

We model the payment card association as a two-sided platform which organises the interactions between a monopolistic Issuer and a monopolistic Acquirer. In our setting, there are two merchants that are differentiated à la Hotelling, and positioned exogenously at the two extremes of a linear city of length one. One merchant provides a good of higher quality than the other. Merchants are homogenous as to their card acceptance benefit and accept payment cards if the merchant fee is not too high. The merchant which produces the good of higher quality can choose to issue its private card. The private card cannot be used by consumers to pay at the other merchant's shop. To issue the private card, the merchant has to pay a fixed cost. We assume that consumers differ across their card usage benefit, which is the same for a given consumer if he pays by card or if he uses the private card.

We start by showing that, in our setting, if the merchant decides to issue its private card, he chooses a transaction fee equal to zero, such that, if consumers come to his shop, they always prefer the private card to the bank-issued payment card. The intuition is that the merchant is active on the market for card transactions and on the product market, and that his incentives on each of these two markets are to set a very low price for the private card. On the market for payment transactions, we show that the merchant has an incentive to undercut the price that

[3] For a review of the literature, see Rochet (2003).

is set by the issuer for the payment card. Also, the merchant chooses a low price for the private card because he obtains a higher benefit per transaction if his consumers pay with the private card than if they pay cash. On the product market, the merchant has an incentive to set a low price for the private card, because he obtains a higher market share, by stealing consumers from his competitor.

We prove that, if one merchant issues private cards, the other merchant becomes less resistant to card acceptance than in the benchmark case, in which no merchant issues private cards. The threat of losing consumers on the product market raises the maximum merchant fee that he is willing to pay to accept payment cards. Since the monopolistic Acquirer chooses the maximum merchant fee compatible with merchants' acceptance of payment cards, the effect of the private card is to increase the merchant fee. On the other hand, the Issuer charges a lower card fee, to compete with the very aggressive price that is set by the merchant for the private card. Therefore, the competition with the private card changes the structure of prices.

Then, we derive the impact of the interchange fee on the merchant's incentives to enter the market for payment card transactions by issuing private cards. We show that there are two effects. A higher interchange fee reduces the card fee, which toughens the competition with the Issuer. This effect lowers the merchant's incentives to issue private cards. On the other hand, a higher interchange fee increases the merchant fee, and hence the costs of the rival merchant, which raises the benefits of issuing private cards. In our setting, the first effect is dominant. Therefore, if entry is not blockaded, the threat of the private card may lead the payment platform to increase its interchange fee so as to deter the merchant from entering the market for payment card transactions. In our model, the payment system always prefers deterring than accommodating entry, if this strategy is feasible. Also, simulations suggest that, if entry is accommodated, the interchange fee is lower than in the benchmark case.

We also consider perfect competition instead of a monopoly on the acquisition side. With perfect competition on the acquisition side, we show that the payment system may set a low interchange fee to deter entry.

The rest of this paper is organised as follows. In section two, we start by presenting the model and the assumptions. In section three, we solve for the equilibrium of the game. In section four, we extend our model by assuming perfect competition on the acquisition side. Finally, we conclude.

2 The model

Two merchants are located at the two extremes of a linear street of length one. Consumers can always pay cash when they decide to buy a good from one of the merchants. There are also two banks, an Issuer and an Acquirer, which provide payment card services to consumers and merchants, respectively. Both banks build a payment card association, which chooses an interchange fee, so as to maximise banks' joint profits. However, the largest merchant may choose to issue its own payment card, the "private card", and compete with the Issuer to provide payment services to its consumers, while bypassing the Acquirer's services.

Our model studies the conditions under which the largest merchant issues a private card, and the optimal reaction of the payment card association.

Merchants: Two merchants, denoted by 1 and 2, are located at the extremities of a linear city of length one. Merchant 1's shop is located at point 0 and merchant 2's shop at point 1. Each merchant i, for $i \in \{1;2\}$, sells a good of quality q_i at a price p_i. We assume that $q_1 > q_2$, and we refer to merchant 1 as the "largest" merchant.[4] We denote $\Delta q = q_1 - q_2$. The marginal costs are the same and equal to c.

We assume that merchant 1 can issue private cards at a fixed cost F. If a consumer decides to use the private card when he goes to merchant 1's shop, he has to pay a transaction fee, f^{PC}. The merchant incurs a cost c_M for each transaction paid by the private card. This cost corresponds to the costs of issuing and acquiring a transaction. We assume that merchant 2 cannot issue a private card.

We also suppose that merchants are homogeneous as regards to their card acceptance benefit, which we denote by b_S, with $b_S \geq 0$. Merchant 1 obtains the same card acceptance benefit whether the transaction is paid by a bank's card or the private card.[5]

Consumers: Consumers are uniformly located along the linear city. They incur a linear transportation cost t when they travel to shop either at merchant 1's or merchant 2's shop. When it decides to shop at merchant i's, each consumer purchases zero or one unit of the good.

In his wallet, each consumer always holds cash and a payment card issued by his bank.[6] He can always use cash at no cost[7] to pay for his expenses. If he decides to use a payment card,

[4] Indeed, in equilibrium, given that the two merchants offer the same payment possibilities, merchant 1 obtains a higher market share than its rival, due to its quality advantage.

[5] In our model, we choose to focus on the bypass decision of the largest merchant. This is why we assume that the private card is a perfect substitute to the payment card, for consumers as well as merchant 1.

[6] In the model, we consider cardholding decisions as exogenous, and focus on the choice of the payment instrument at the point of sales.

[7] The costs and the benefits of using cash are normalised to zero.

he has to pay a transaction fee to the issuer of the card. The payment card issued by the bank can be used either to buy from merchant 1 or merchant 2, provided it is accepted at the point of sales. A consumer may also hold a private card, issued by merchant 1, which can only be used to purchase a good at merchant 1's shop.

Each consumer is characterised by his benefit, b_B, of using a card rather than cash. We assume that the benefit b_B is the same whether the card is issued by the bank or by merchant 1, and that b_B is uniformly distributed over $[0, 1]$. One interpretation is that they may attach different values to the convenience of using a card rather than cash.

A consumer located at x, whose card usage benefit is b_B, and who buys from merchant i located at x_i, enjoys a net utility of:

$$U = v + t\,|x - x_i| + q_i - p_i + b_B - f,$$

if he uses his card, and pays the transaction fee f, and a net utility of

$$U = v + t\,|x - x_i| + q_i - p_i,$$

if he pays cash, where v represents a fixed utility obtained from consuming the good. We assume that v is sufficiently high such that the market is covered.

Banks: The Issuer (I) and the Acquirer (A) are monopolists.[8] For each transaction, the Issuer charges card-users with a fee, f^C, and the Acquirer charges merchants with a fee, $m \geq 0$. The Acquirer pays to the Issuer a per-transaction interchange fee, denoted by a^P, with $a^P \geq 0$. Banks' have constant marginal costs c_i per transaction, for $i = I, A$, and profits are denoted by Π_I and Π_A. If no merchant accepts cards, banks make no profits, i.e., $\Pi_i = 0$ for $i = I, A$.

Payment system: The payment system (S) chooses the interchange fee, a^P, which maximises the sum of banks' profits, $\Pi_S = \Pi_I + \Pi_A$. We assume that the Non-Discrimination Rule (NDR) holds, which means that merchants are forbidden to charge different prices according to the payment instrument used for the transaction.

Finally, we define the social welfare, W, as the sum of consumers' surplus, S_C, merchants' surplus, S_M, and banks' profits, $\Pi_I + \Pi_A$. We also make the following assumptions.

Assumption 1. $t \geq \Delta q/3 + 11/3$.

[8] In Section 4, we will discuss how the market structure on the acquisition side affects our results.

This assumption ensures that Π_I is concave with respect to f^C.[9]

Assumption 2. $q_i \in [0,1]$ for $i \in \{1;2\}$ and $q_1 > q_2$.

Assumption 3. $0 \leq c_M \leq b_S \leq c_I + c_A < 1$

The fact that $b_S \geq c_M$ implies that merchant 1 makes a net benefit for each transaction paid with the private card. We also assume that $c_M \leq c_I + c_A$, which means that merchant 1 is at least as efficient as the association of the Issuer and the Acquirer. Finally, since $b_S \leq c_I + c_A < 1$, it is socially optimal that some consumers but not all pay with their payment cards.

Timing: The timing of the game is as follows:

1. The payment platform chooses the interchange fee, a^P, which maximises the joint profits of the banks.
2. Merchant 1 decides whether or not to issue a private card.
3. Banks choose simultaneously and non-cooperatively their transaction fees, f^C and m, and merchant 1 decides simultaneously on the private card transaction fee, f^{PC}.
4. Merchants choose their prices p_1 and p_2, and whether or not to accept cards.
5. Consumers decide which payment instrument to use (cash, payment card or private card), and which merchant to buy from.

With this timing, we assume that merchant 1 decides whether or not to issue a private card, once the interchange fee has been set. Indeed, in practice, payment platforms do not adjust the level of the interchange fee very frequently. Besides, we choose to focus on the effect of the interchange fee on the merchant's incentives to bypass the payment system.

We look for the subgame perfect equilibrium, and solve the game by backward induction.

3 A benchmark: no private card

We start by analysing a benchmark, in which we assume that it is too costly for merchant 1 to issue private cards.[10] We determine the condition under which both merchants accept payment cards, and the optimal interchange fee chosen by the payment card system.

[9] See Appendix D1.

[10] This benchmark also corresponds to the subgame in which merchant 1 does not issue a private card.

This benchmark case is close to Rochet and Tirole (2002). But, in our setting, we assume that banks on each side of the payment platform are monopoly. Whereas, in Rochet and Tirole, there is perfect competition in the acquisition market and imperfect competition in the issuing market.

We focus on the equilibrium in which both merchants accept cards.[11] Let $(a^P)^B$, $(f^C)^B$ and $(m)^B$ denote the equilibrium interchange fee, transaction fee and merchant fee, respectively. We denote by $\pi_i^B\left((f^C)^B, m^B\right)$ the equilibrium profit of merchant i.

Proposition 1 *If merchant* 1 *cannot issue private cards, both merchants accept payment cards if*

$$m \leq b_S + \frac{1 - f^C}{2}.$$

The optimal interchange fee is

$$(a^P)^B = 2(b_S - c_A) + 1 - c_I,$$

and the optimal transaction fees are

$$\begin{aligned} (f^C)^B &= c_I + c_A - b_S, \\ (m)^B &= \frac{3b_S + 1 - c_I - c_A}{2}. \end{aligned}$$

Proof. See Appendix A. ■

As in Rochet and Tirole (2002, 2006), we find that strategic merchants are ready to pay for a higher merchant fee, to attract consumers to their stores. They internalise a fraction of the cardholders' benefit of using their cards. If merchants were not strategic, the maximum merchant fee compatible with card acceptance would be b_S, and the optimal interchange fee would be $a^P = b_S - c_A$.

4 The equilibrium with private cards

In this Section, we assume that merchant 1 can issue private cards[12] and we determine the equilibrium of the game, starting from the last stage.

[11] There might also be "high resistance" equilibria as in Rochet and Tirole (2002), in which no merchant accepts cards.

[12] Or, equivalently, that the fixed cost of a private card system is not prohibitively high.

4.1 Stage 5 and 4: card acceptance decisions and prices

If merchant 1 does not issue private cards, the analysis is similar to the benchmark case. From now on, we assume that merchant 1 issues private cards, and we determine the demands for merchants 1 and 2. We denote $\Delta f = f^C - f^{PC}$ and we assume that consumers who shop at merchant 1's and are indifferent between the payment card and the private card use the private card.

At stage 5, consumers take into account the price and the quality of the good in their decision to shop either at merchant 1's or merchant 2's, as well as the availability of each payment instrument. As each merchant can either accept or refuse cards, we have four possible cases, depending on the merchants' acceptance decisions. We denote by $\pi_i^{x_1,x_2}$ the profit of merchant i, where x_i denotes the card acceptance decision of merchant i. We set $x_i = NC$ if merchant i refuses payment cards and $x_i = C$ if he accepts cards. At stage 4, merchant i chooses the price p_i that maximises his profit,

$$\pi_i^{x_1,x_2} = \left(D_i^{PC} + D_i^C + D_i^{Cash}\right)(p_i - c) + (f^{PC} + b_S - c_M)D_i^{PC} + (b_S - m)D_i^C,$$

where D_i^{PC}, D_i^C, and D_i^{Cash} denote the demand of consumers who shop at merchant i's and pay with the private card, the payment card and cash, respectively. Notice that $D_2^{PC} = 0$ as, by assumption, merchant 2 does not issue private cards.

We determine below the equilibrium of stages 4 and 5 in each of the four possible cases, for $(x_1, x_2) \in \{NC, C\}^2$.

4.1.1 Both merchants accept payment cards

We start by analyzing consumers' decisions at stage 5. Since both merchants accept payment cards, consumers trade off between the private card, the payment card and cash when they shop at merchant 1's and trade off between the payment card and cash when they shop at merchant 2's. If $f^{PC} > f^C$, consumers who shop at merchant 1's always use their payment card instead of the private card, as their net utility from using the payment card, $b_B - f^C$, is strictly greater than their net utility of using the private card, $b_B - f^{PC}$. Therefore, the demands for merchant 1 and merchant 2 are identical to their demands in the benchmark case, if they both accept payment cards, and can be found in Appendix A.

If $f^{PC} \leq f^C$, consumers who shop at merchant 1's prefer the private card to the payment card. When they trade off between merchant 1 and merchant 2, consumers take into account both the net utility associated to the product purchase and the net utility that they obtain

from the payment transaction. Consumers such that $b_B < f^{PC} \leq f^C$ always pay cash, as their net utility from a payment by card is negative. A standard Hotelling analysis shows that each merchant i obtains a share w_i of these consumers, where

$$w_i = \frac{1}{2} + \frac{1}{2t}(q_i - q_j + p_j - p_i),$$

for $(i;j) \in \{1;2\}^2$ and $i \neq j$. By integrating for $b_B \in \left[0, f^{PC}\right]$, we obtain that the demand from cash users is equal to $f^{PC} w_i$ for merchant i.

Consumers such that $b_B \in \left[f^{PC}, f^C\right)$ trade off between purchasing from merchant 1 and paying with the private card and purchasing from merchant 2 and paying cash, as their net utility from a payment by card $(b_B - f^C)$ is negative, whereas their net utility from a payment with the private card $(b_B - f^{PC})$ is positive. The marginal consumer is given by

$$v - p_1 - tx + q_1 + b_B - f^{PC} = v - p_2 - t\,(1 - x) + q_2,$$

that is,

$$x\,(b_B) = \frac{\Delta q + p_2 - p_1 + b_B - f^{PC}}{2t}.$$

Aggregating for $b_B \in \left[f^{PC}, f^C\right)$, the demand for merchant 1 from these consumers is

$$\int_{f^{PC}}^{f^C} x\,(b_B)\,db_B = \frac{p_2 - p_1 + \Delta q}{2t}\Delta f + \frac{(\Delta f)^2}{4t},$$

whereas the demand for merchant 2 from these consumers is

$$\int_{f^{PC}}^{f^C} (1 - x\,(b_B))\,db_B = \frac{p_1 - p_2 - \Delta q}{2t}\Delta f + \Delta f - \frac{(\Delta f)^2}{4t}.$$

Consumers such that $b_B \geq f^C$ trade off between purchasing from merchant 1 and paying with the private card and purchasing from merchant 2 and paying with the payment card, since their net utility of paying by card is positive. The marginal consumer is given by

$$v - p_1 - tx + q_1 + b_B - f^{PC} = v - p_2 - t\,(1 - x) + q_2 + b_B - f^C,$$

that is,

$$x = \frac{\Delta q + p_2 - p_1 + \Delta f}{2t},$$

therefore, aggregating over $b_B \in \left[f^C, 1\right]$, the demand for merchant 1 from these consumers is

$$\left(1 - f^C\right) \frac{\Delta q + p_2 - p_1 + \Delta f}{2t},$$

whereas the demand for merchant 2 from these consumers is

$$\left(1 - f^C\right) \left[1 - \frac{\Delta q + p_2 - p_1 + \Delta f}{2t}\right].$$

To sum up, we find that the demand of cash users is

$$D_1^{Cash} = f^{PC} w_1,$$

and

$$D_2^{Cash} = f^C w_2 - \frac{(\Delta f)^2}{4t},$$

for merchant 1 and merchant 2, respectively. Compared to the benchmark case, the demand of cash users for merchant 1 is determined by the price of the private card, which plays the same role as the payment card. If the price of the private card is lower than the price of the payment card, merchant 2 loses some of its cash users who prefer to shop at merchant 1's and pay with the private card. This corresponds to the second term in D_2^{Cash}.

The demand of card users is the demand of private card users for merchant 1,

$$D_1^{PC} = (1 - f^{PC}) w_1 + \frac{(1 - f^C)\Delta f}{2t} + \frac{(\Delta f)^2}{4t}, \tag{1}$$

and the demand of payment card users for merchant 2,

$$D_2^C = (1 - f^C) w_2 - \frac{(1 - f^C)\Delta f}{2t}. \tag{2}$$

If the private card is less expensive than the payment card, merchant 1 attracts some cash users and some card users from merchant 2. The number of cash users who switch from merchant 2 to merchant 1 is given by the third term in (1), that is, $(\Delta f)^2 / (4t)$. The number of card users who switch from merchant 2 to merchant 1 is given by the second term in (1), that is, $(1 - f^C)\Delta f / (2t)$.

We now turn to stage 4 of our game. Merchant 1 makes profit

$$\pi_1^{C,C} = \left(D_1^{Cash} + D_1^{PC}\right)(p_1 - c) + (b_S + f^{PC} - c_M) D_1^{PC},$$

whereas merchant 2 makes profit

$$\pi_2^{C,C} = \left(D_2^{Cash} + D_2^C\right)(p_2 - c) + (b_S - m)D_2^C.$$

When merchants decide on their prices, they take into account both their net revenues from product sales (the first term in the profit functions) and the costs or benefits associated to card payments (the second term in the profit functions). Replacing for the expressions of demands in π_1 and π_2, and solving for the first order conditions,[13] we find that the equilibrium prices are

$$p_1 = c + t + \frac{\Delta q}{3} + \frac{1}{3}\left(\frac{(\Delta f)^2}{2} + (\Delta f)(1 - f^C) + (m - b_S)(1 - f^C) - 2\left(f^{PC} + b_S - c_M\right)(1 - f^{PC})\right),$$

and

$$p_2 = c + t - \frac{\Delta q}{3} - b_S + \frac{1}{3}\left(-\frac{(\Delta f)^2}{2} - (\Delta f)(1 - f^C) + 2(m - b_S)(1 - f^C) - \left(f^{PC} + b_S - c_M\right)(1 - f^{PC})\right).$$

A higher fee for the private card has two opposite effects on equilibrium prices. First, a higher f^{PC} decreases merchant 1's perceived marginal cost for the transactions paid by the private card, which tends to reduce merchants' prices. Second, a higher f^{PC} reduces the volume of transactions paid by the private card, hence, leads to a higher *average* perceived marginal cost for merchant 1. This is because the perceived marginal cost for transactions paid cash, c, is higher than the perceived marginal cost for transactions paid by the private card, $c - \left(b_S + f^{PC} - c_M\right)$, since $b_S > c_M$. For sufficiently low values of f^{PC}, the first effect dominates the second effect, and prices decrease with the private card fee. On the contrary, for sufficiently high values of f^{PC}, prices increase with the private card fee.

Replacing for the equilibrium values of p_1 and p_2 in $\pi_1^{C,C}$ and $\pi_2^{C,C}$, we obtain the equilibrium profits, which can be found in Appendix B.

4.1.2 Merchant 1 does not accept payment cards, while merchant 2 accepts them

If $f^{PC} \leq f^C$, whether merchant 1 accepts payment cards or not, his card consumers will always use the private card as it is cheaper. Therefore, whether he accepts cards or not, merchant 1 will face the same demand, and the equilibrium prices and profits are identical to the previous case.

If $f^{PC} > f^C$, a similar analysis as in the previous section shows that the demand of cash

[13] The second order condition is verified.

users is

$$D_1^{Cash} = f^{PC} w_1 - \frac{(\Delta f)^2}{4t},$$

and

$$D_2^{Cash} = f^C w_2,$$

for merchant 1 and merchant 2, respectively. The second term in D_1^{Cash} represents the cash consumers of merchant 1 who decide to purchase from merchant 2 and pay by card. The demand of card users is the demand of private card users for merchant 1,

$$D_1^{PC} = (1 - f^{PC}) w_1 + \frac{(1 - f^{PC})\Delta f}{2t},$$

and the demand of payment card users for merchant 2,

$$D_2^C = (1 - f^C) w_2 - \frac{(1 - f^{PC})\Delta f}{2t} + \frac{(\Delta f)^2}{4t}.$$

The second term in D_1^{PC} is negative (as $\Delta f < 0$) and represents the private card users who prefer to shop at merchant 2's and pay with the payment card. This term corresponds to the second term in D_2^C. The last term in D_2^C corresponds to the cash consumers of merchant 1 who decide to shop at merchant 2's and pay by card.

The equilibrium prices are

$$p_1 = c + t + \frac{\Delta q}{3} + \frac{1}{3}\left(-\frac{(\Delta f)^2}{2} + (\Delta f)(1 - f^{PC}) + (m - b_S)(1 - f^C) - 2(f^{PC} + b_S - c_M)(1 - f^{PC}) \right),$$

and

$$p_2 = c + t - \frac{\Delta q}{3} + \frac{1}{3}\left(\frac{(\Delta f)^2}{2} + -(\Delta f)(1 - f^{PC}) + 2(m - b_S)(1 - f^C) - (f^{PC} + b_S - c_M)(1 - f^{PC}) \right).$$

The effect of f^{PC} on prices is similar to the previous case. Equilibrium profits, $\pi_i^{NC,C}$, can be found in Appendix B.

4.1.3 Merchant 1 accepts all payment cards, while merchant 2 refuses them

If $f^{PC} > f^C$, private cards are never used by consumers. This case is identical to the benchmark case, in which merchant 2 does not accept cards, while merchant 1 accepts them.

If $f^{PC} \leq f^C$, payment cards are never used by consumers, as merchant 2 does not accept cards, and consumers prefer to use the private card when they shop at merchant 1's. Using the same analysis as in the previous cases, we find that the demands of cash users for merchant 1

and merchant 2 are

$$D_1^{Cash} = f^{PC} w_1,$$

and

$$D_2^{Cash} = w_2 - \frac{(1 - f^{PC})^2}{4t},$$

respectively. The demand of private card users for merchant 1 is

$$D_1^{PC} = (1 - f^{PC}) w_1 + \frac{(1 - f^{PC})^2}{4t}.$$

The second term in D_1^{PC} represents the cash users of merchant 2 who decide to shop at merchant 1's and pay with the private card.

The equilibrium prices are

$$p_1 = c + t + \frac{\Delta q}{3} + \frac{1}{3}\left(\frac{(1 - f^{PC})^2}{2} - 2(f^{PC} + b_S - c_M)(1 - f^{PC})\right),$$

and

$$p_2 = c + t - \frac{\Delta q}{3} + \frac{1}{3}\left(-\frac{(1 - f^{PC})^2}{2} - (f^{PC} + b_S - c_M)(1 - f^{PC})\right),$$

and the equilibrium profits, $\pi_i^{C,NC}$, can be found in Appendix B. The effect of f^{PC} on prices is similar to the previous cases.

4.1.4 Both merchants refuse payment cards

As consumers trade off between the private card and cash at merchant 1's and can only pay cash at merchant 2's, the demands are identical to the previous case in which merchant 2 refuses cards but not merchant 1, and $f^{PC} \leq f^C$. Equilibrium prices and equilibrium profits, $\pi_i^{NC,NC}$, are also identical.

4.1.5 Card acceptance conditions

At stage 4, simultaneously with setting prices, the merchants decide whether or not to accept cards. The situation in which both merchants accept cards constitutes a Nash equilibrium if and only if

$$\pi_1^{C,C}\left(m, f^C, f^{PC}\right) \geq \pi_1^{NC,C}\left(m, f^C, f^{PC}\right),$$

and

$$\pi_2^{C,C}\left(m, f^C, f^{PC}\right) \geq \pi_2^{C,NC}\left(m, f^C, f^{PC}\right).$$

The first condition means that merchant 1 has no incentive to deviate to the equilibrium in which merchant 2 is the only one who accepts cards. The second condition means that merchant 2 makes more profit if both merchants accept cards than in a situation where merchant 1 is the only one who accepts cards. The card acceptance decisions depend on the transaction fees, m, f^C and f^{PC}, which are set at stage 3 of the game.

4.2 Stage 3: choice of transaction fees

In this section, we assume that merchant 1 issues private cards, and we determine the transaction fees chosen by the banks and merchant 1.[14] We show that there exists an equilibrium in which both merchants accept payment cards, and that in this equilibrium, merchant 1 sets $f^{PC} = 0$.

We start by analyzing the decision of merchant 1. For given m and f^C, merchant 1 chooses the private card fee, f^{PC}, so as to maximise his profit,

$$\pi_1^{x_1,x_2} = \left(D_1^{PC} + D_1^C + D_1^{Cash}\right)(p_1 - c) + (f^{PC} + b_S - c_M)D_1^{PC} + (b_S - m)\,D_1^C. \qquad (3)$$

The following proposition shows that merchant 1's best response has a remarkable property.

Proposition 2 *If merchant* 1 *issues the private card, for any* m *and* f^C*, his best response is to choose a transaction fee equal to zero, that is,* $f^{PC} = 0$.

Proof. In Appendix C1, we show that if $f^{PC} < f^C$, merchant 1's profit decreases with the price of the private card, f^{PC}. Consequently, in this case, for any m, his best response is to set $f^{PC} = 0$. In Appendix C2, we show that merchant 1 always makes more profit if he undercuts f^C by choosing $f^{PC} < f^C$. Therefore, merchant 1's best response is to choose a transaction fee which is equal to zero. ■

Proposition 2 shows that merchant 1 sets a very aggressive price for his private card, which is below cost. The intuition is that merchant 1 is active on two markets, the market for card transactions and the product market, and that his incentives in each of these two markets are to set a low fee for the private card.

On the market for card transactions, merchant 1 competes in prices with bank I. Since the payment card and the private card are perfect substitutes, if f^C is sufficiently high, then the Bertrand logic applies, and merchant 1 has an incentive to undercut the price of the payment card. Indeed, from the analysis of stage 4 and 5, we know that if merchant 1 accepts payment cards and if he undercuts the Issuer by setting a slightly lower fee, that is, $f^{PC} = f^C - \epsilon$,

[14] If merchant 1 does not issue private cards, this is the benchmark case, that we have analyzed in Section 3.

with ϵ small, then the demand of card payments (either with a payment or a private card) remains unchanged. Consequently, merchant 1 has an incentive to undercut bank I if $f^C + b_S - c_M \geq b_S - m$ (see term (II) in (3)). Apart from this competitive effect on the market for card transactions, merchant 1 also has an incentive to lower his private card fee to encourage consumers to pay with the private card instead of cash, as he earns a higher benefit with the private card.

The private card fee has also an impact on competition in the product market. First, merchant 1 has an incentive to set a low fee to attract consumers of merchant 2, which prefer to shop at merchant 1's for lower transaction costs. Second, a lower f^{PC} softens competition on the product market because it increases the perceived marginal cost of merchant 1 for card transactions.

The effects of f^{PC} on the profits made on the product market and the market for payment transactions go in the same direction, and provide merchant 1 with strong incentives to set a very low private card fee.

We have proved that $f^{PC} = 0$ constitutes a dominant strategy for merchant 1. Therefore, from this point, we analyse the decisions of bank I and bank A for $f^{PC} = 0$.

As $f^{PC} = 0$, for any f^C and m, the consumers of merchant 1 always prefer the private card to the payment card. Hence, the payment card may only be used by consumers of merchant 2. If merchant 2 refuses the payment card, banks do not make any profit. Therefore, bank I and A choose f^C and m, under the constraint that merchant 2 accepts cards. We show that this is the case for sufficiently low values of m.

Lemma 1 *There exists* $\widetilde{m}(f^C) \in \big(b_S + (1 - f^C)/2; b_S + 3(1 - f^C)/4\big)$*, such that merchant* 2 *accept payment cards for* $m \leq \widetilde{m}(f^C)$*. Merchant* 1 *is indifferent between accepting and refusing payment cards.*

Proof. See Appendix D. ■

As in the benchmark case, if the merchant fee is sufficiently low, there is an equilibrium in which both merchants accept payment cards. Since $f^{PC} = 0$, consumers always choose the private card to pay at merchant 1's. Hence, merchant 1 is indifferent between accepting and refusing payment cards. Merchant 2 accepts payment cards for sufficiently low values of m.

Corollary 1 *For a given payment card fee,* f^C*, the merchants are less resistant to card acceptance if merchant* 1 *issues private cards than if it does not.*

Proof. Indeed, in the benchmark case, the card acceptance condition was

$$m \leq b_S + \frac{1 - f^C}{2},$$

while we have

$$\widetilde{m}(f^C) \geq b_S + \frac{1 - f^C}{2}.$$

■

Merchant 2's incentive to deviate from the equilibrium in which both merchants accept cards is equal to the difference between his profit in case of deviation and his statu-quo profit.

In the benchmark case, as in Rochet and Tirole (2002), merchant 2's decision to refuse cards has two effects on his profit, a "perceived marginal cost effect", and a "market share effect". First, merchants' perceived marginal costs change if merchant 2 refuses cards, as he saves the merchant fee, m, net of the benefit of being paid by card, b_S. Therefore, his perceived marginal cost decreases if $m - b_S > 0$, and increases otherwise. Besides, when merchant 2 refuses cards, the proportion of card users at merchant 1's increases. Hence, merchant 1's average perceived marginal cost increases if $m - b_S > 0$, and decreases otherwise. Consequently, the higher m, the higher the benefits of deviation for merchant 2. Second, if he decides to refuse cards, merchant 2 may lose market share, as some of his card users may decide to switch to merchant 1. This market share effect makes deviation less profitable for merchant 2. Its magnitude is higher when the payment card fee is lower.

If merchant 1 issues a private card, merchant 2's incentives to refuse cards also depend on a perceived marginal cost effect and a market share effect. The market share effect is comparable to the one observed in the benchmark case, except that its magnitude is higher because merchant 1 sets a private card fee equal to zero. This reduces merchant 2's incentives to deviate, in comparison to the benchmark case. The perceived marginal cost effect has a different impact on merchant 1, since consumers always prefer the private card to the payment card when they shop at merchant 1's. For merchant 1, the perceived marginal cost of private card payments is negative (equal to $c_M - b_S$). Hence, when merchant 2 deviates and refuses cards, the proportion of private card users at merchant 1's increases, which reduces the average perceived marginal cost of merchant 1 (even if $m > b_S$). Therefore, merchant 2's incentives to deviate are lower compared to the benchmark case.

This explains why merchant 2 is less resistant to card acceptance. A direct consequence is that, for a given f^C, the Acquirer can set a higher merchant fee if merchant 1 issues private cards.

We now determine the transaction fees, f^C and m, that maximise the profits of the Issuer and the Acquirer, respectively, for $m \leq \widetilde{m}(f^C)$, that is,

$$\Pi_I = \left(f^C + a^P - c_I\right) D_2^C,$$

and

$$\Pi_A = (m - a^P - c_A) D_2^C,$$

where, from Appendix C,

$$D_2^C = \frac{1}{2t}(1 - f^C)\left[t + \frac{1}{3}\left(-\Delta q - (b_S + 1 + f^C)f^C + c_M - m(1 - f^C)\right)\right]. \tag{4}$$

The Issuer and the Acquirer trade off between a higher margin and a higher volume of card transactions. Notice, from equation (4), that the volume of card transactions is decreasing with the merchant fee, m. This is because the merchant fee is passed to consumers through merchant 2's perceived marginal cost. Solving for the first order conditions yields[15]

$$\frac{d\Pi_A}{dm} = \frac{dD_2^C}{dm}(m - a^P - c_A) + D_2^C = 0, \tag{5}$$

and

$$\frac{d\Pi_I}{df^C} = \frac{dD_2^C}{df^C}(f^C + a^P - c_I) + D_2^C = 0. \tag{6}$$

The following proposition shows that there exists a unique equilibrium in which both merchants accept cards, and that the Acquirer chooses the maximum merchant fee compatible with merchant acceptance.

Proposition 3 *There exists a unique equilibrium, such that merchant* 1 *sets* $f^{PC} = 0$*, the Acquirer chooses the maximum merchant fee compatible with merchant* 2*'s card acceptance, and the issuer chooses a strictly positive card fee.*

Proof. See Appendix E. ■

The optimal merchant fee must be compatible with merchant 2's non deviation condition, as the Acquirer makes zero profit if merchant 2 deviates from the equilibrium in which the merchants accept cards. In Appendix D, we show that the merchant fee that maximises the Acquirer's profit does not satisfy the non deviation condition. Hence, since Π_A is concave in m,[16] the optimal merchant fee is equal to $\widetilde{m}(f^C)$.

[15] In Appendix E-2, we prove that the second order conditions are verified if t is sufficiently high.
[16] This is proved in Appendix D.1.

Proposition 4 *The merchant fee is higher in the presence of a private card, while the transaction fee chosen by the Issuer for the payment card is lower, that is, we have* $m^* > m^B$ *and* $(f^C)^* < (f^C)^B$.

Proof. See Appendix F. ■

When merchant 1 issues private cards and sets a very aggressive private card fee, the Issuer reacts by setting a lower payment card fee than in the benchmark case. Notice, however, that the Issuer's reaction cannot be explained only by the competition with the private card on the market for payment transactions, since f^{PC} is set to zero and $(f^C)^* > 0$.

The Issuer's reaction is also related to the product market. By setting $f^{PC} = 0$, merchant 1 obtains what could be interpreted as a quality advantage over merchant 2, which reduces the demand of merchant 2, including the demand from card users. The Issuer has an incentive to reduce merchant 2's quality disadvantage by lowering the payment card fee; in other words, the Issuer internalises the effect of the payment card fee on competition in the product market. As the payment card fee is reduced, the Acquirer can increase his merchant fee, since $\widetilde{m}(f^C)$ is decreasing in f^C.

Hence, the effect of the private card is to reinforce the market power of the Acquirer, as it makes merchants less resistant to card acceptance. On the contrary, the private card reduces the market power of the Issuer, because the latter has to lower the payment card fee to stimulate the demand of card users at merchant 2's. A consequence is that the price structure of the payment platform changes because of the competition with the private card. To analyse the effect of the introduction of private cards on the total price that is charged by the payment platform, $f^C + m$, we have to revert to numerical simulations. They suggest that the total price is lower if the payment platform faces the competition of the private card.

4.3 Stage 2: decision to issue a private card

Merchant 1 decides to issue private cards if and only if

$$\pi_1^{C,C}\left((f^{PC})^*, (f^C)^*, m^*\right) - F \geq \pi_1^B\left((f^C)^B, m^B\right). \tag{7}$$

Notice that this corresponds to a vertical integration decision, except that it takes place in a two-sided market, that is, merchant 1 has to decide whether or not to create his own payment platform.

4.4 Stage 1: choice of the interchange fee

In this section, we start by conducting some comparative statics with respect to the interchange fee, if merchant 1 issues private cards. Then, we determine the optimal level of the interchange fee. We compare the optimal interchange fee with the one obtained in the benchmark case.

Comparative statics We assume that merchant 1 issues private cards, that is, condition (7) is satisfied. We analyse the effect of the interchange fee on the optimal transaction fees chosen by the Issuer and the Acquirer.

Lemma 2 *The transaction fee chosen by the Issuer for the payment card is decreasing with a^P, while the merchant fee chosen by the Acquirer is increasing with a^P, that is, we have $d(f^C)^*/da^P < 0$ and $d(m)^*/da^P > 0$.*

Proof. See Appendix G. ■

The interchange fee, a^P, has a direct and a strategic effect on the transaction fees, f^C and m. First, a higher a^P implies a lower perceived marginal cost for bank I and a higher perceived marginal cost for bank A. Therefore, bank I has incentives to decrease f^C, while bank A is willing to increase m. Second, we show in Appendix D3 that m and f^C are strategic substitutes. Therefore, a higher a^P implies a lower f^C, which in turn implies a higher m. Similarly, a higher a^P implies a higher m, hence a lower f^C. As the direct effect and the strategic effect have the same sign, we find that the payment card fee decreases with a^P, whereas the merchant fee increases with a^P.

We now study the impact of the interchange fee on entry. The entry condition, given by (7), can be rewritten as $EC(a^P) \geq 0$, where

$$EC(a^P) = \Psi^2 - 2tF - \left(t + \frac{\Delta q}{3}\right)^2,$$

and $\Psi = t + \frac{1}{3}(\Delta q - c_M + \left((f^C)^* + m^*\right)(1 - (f^C)^*) + \frac{((f^C)^*)^2}{2} + b_S(f^C)^*)$. Taking the derivative of EC with respect to a^P, we obtain

$$(EC)'(a^P) = \frac{2}{3}\Psi \times \left[\underbrace{(b_S + 1 - (f^C)^* - m^*)\frac{d(f^C)^*(a^P)}{da^P}}_{\text{(I)}} + \underbrace{(1 - (f^C)^*)\frac{dm^*(a^P)}{da^P}}_{\text{(II)}}\right].$$

From assumption 1, we have $\Psi \geq 0$. Since $b_S + 1 - (f^C)^* - m^* > 0$ from Lemma 1, and $d(f^C)^*/da^P < 0$ from Lemma 2, then term (I) is negative. Term (II) is positive as $dm^*/da^P > 0$,

from Lemma 2. This shows that the interchange fee impacts merchant 1's incentives to issue private cards in two opposite ways. If the interchange fee increases, the perceived marginal cost of merchant 2 rises through the payment of the merchant fee. Therefore, merchant 1 benefits from a reduction of the demand of merchant 2, since the latter is forced to increase its price. Merchant 1's incentives to issue its private card become higher, because it saves him the cost of the merchant fee, which has increased, while giving him the opportunity of increasing its market share. At the same time, if the interchange fee increases, this triggers a reduction of the payment card transaction fee, which yields a higher demand for merchant 2, and lowers the incentives of merchant 1 to issue its payment card.

The following Lemma shows that the first effect always dominates the second effect, that is, $EC(a^P)$ is decreasing with a^P.

Lemma 3 *A higher a^P reduces merchant 1's incentives to issue private cards.*

Proof. See Appendix I. ■

Optimal interchange fee Now, we determine the interchange fee that maximises banks' joint profits, that we denote by $\left(a^P\right)^*$.

From Lemma 2, we know that when the interchange fee increases, the Acquirer charges a higher merchant fee. However, from Lemma 1, the equilibrium merchant fee is bounded from above. Therefore, we can define $a^{\max}$ as the highest value of the interchange fee, a^P, such that the Acquirer's margin is positive, that is, $\widetilde{m}\left(\left(f^C\right)^*\left(a^P\right)\right) - a^P - c_A \geq 0$.

From Lemma 3, we know that $EC(a^P)$ is decreasing in a^P, for $a^P \in [0, a^{\max}]$. Whether merchant 1 issues private cards or not depends in particular on the sign of $EC(a^{\max})$. If $EC(a^{\max}) \geq 0$, then for all $a^P \in [0, a^{\max}]$, we have $EC(a^P) \geq 0$, which means that merchant 1 always issues private cards. If $EC(a^{\max}) < 0$, since $EC\left(a^P\right)$ is decreasing in a^P, there exists $\widehat{a} \in [0, a^{\max}]$ such that $EC(a^P) < 0$ for $a^P > \widehat{a}$, and $EC(a^P) \geq 0$ otherwise. Hence, there is entry for low values of the interchange fee, and no entry for high values of the interchange fee.

Therefore, the banks' joint profits can be written as

$$(\Pi_I + \Pi_A)(a^P) = \begin{cases} ((f^C)^* + m^* - c_I - c_A) D_2^C(a^P) & \text{if} \quad a^P \leq \widehat{a} \\ (\Pi_I + \Pi_A)^B(a^P) & \text{if} \quad a^P > \widehat{a} \end{cases}.$$

If $a^P \leq \widehat{a}$, the payment platform faces the competition of the private card that is issued by merchant 1. Whereas, if $a^P > \widehat{a}$, entry is deterred and the payment platform makes the same profit as in the benchmark case. In the following Proposition, we characterise the possible equilibrium outcomes.

Proposition 5 *The equilibrium can be characterised by either:*

(i) entry accommodation: the payment system cannot deter entry; it chooses the interchange fee that maximises its profit conditional on the fact that merchant 1 *issues private cards.*

(ii) blockaded entry: for any value of the interchange fee, merchant 1 *has no incentive to issue private cards; the payment system sets* $\left(a^P\right)^* = \left(a^P\right)^B$ *and there is no entry.*

(iii) entry deterrence: the payment system sets $\left(a^P\right)^* = \widehat{a}$ *and deters the issuing of the private card; we have* $\widehat{a} > \left(a^P\right)^B$.

Proof. If $EC(a^{\max}) \geq 0$, then for all $a^P \in [0, a^{\max}]$, we have $EC(a^P) \geq 0$. Therefore, the payment system accommodates the entry of merchant 1 on the market for payment transactions.

If $EC(a^{\max}) < 0$, there can be either blockaded entry or entry deterrence. If $\left(a^P\right)^B \geq \widehat{a}$, the payment system can block entry by setting the reference interchange fee, that is, $\left(a^P\right)^B$. Then, provided that the payment system makes more profit by blocking than by accommodating entry, the optimal interchange fee is $\left(a^P\right)^* = \left(a^P\right)^B$, and merchant 1 does not issue private cards.

If $\left(a^P\right)^B < \widehat{a}$, and if the payment system sets $\left(a^P\right)^B$, merchant 1 enters the market for payment transactions. If it benefits from deterring entry, the payment system has to set an interchange fee greater than or equal to $\widehat{a}$. As $\Pi_I + \Pi_A$ is concave in a^P in the benchmark case (See Appendix A), the profit of the payment system is decreasing with a^P for $a^P \geq \left(a^P\right)^B$. Therefore, the optimal interchange fee is the smallest value that blocks entry, that is, $\left(a^P\right)^* = \widehat{a}$.

■

Proposition 5 shows that the threat of the competition with the private card may lead the payment system to *increase* its interchange fee, in comparison to the benchmark case. Indeed, from Lemma 3, we know that if entry is not blockaded, then the payment system has to increase the interchange fee so as to reduce merchant 1's incentives to issue private cards.

When $\widehat{a} \in [0, a^{\max}]$, it remains to determine whether the payment system prefers to accommodate or to deter entry. With the following Proposition, we show that, if Δq is sufficiently high, the payment system always prefer to deter merchant 1 from issuing private cards.

Proposition 6 *Assume that* $\Delta q \geq 1/2$. *Then, if* $\widehat{a} \in [0, a^{\max}]$, *the payment system sets* $\left(a^P\right)^* = \widehat{a} \geq \left(a^P\right)^B$, *and merchant* 1 *does not issue private cards, otherwise, there is entry accommodation.*

Proof. In Appendix J, we prove that $(\Pi_I + \Pi_A)^{PC}(a^P) < (\Pi_I + \Pi_A)^B(\left(a^P\right)^B)$. Therefore, if $\widehat{a} \in [0, a^{\max}]$, the payment system can deter entry and obtains higher profit by doing so. Since $\widehat{a} \geq \left(a^P\right)^B$, we also have that $\left(a^P\right)^* \geq \left(a^P\right)^B$. ■

If Δq is low, simulations suggest that the payment system still prefers deterring to accommodating entry. Also, simulations give the intuition that, with entry accommodation, $\left(a^P\right)^* < \left(a^P\right)^B$.

5 Discussion: Perfect competition between Acquirers

In this section, we discuss how the market structure on the acquisition side impacts the incentives of merchant 1 to issue private cards. So far, we assumed that the payment platform organised the interactions between a monopolistic Issuer and a monopolistic Acquirer. Now, we assume perfect competition on the acquisition side.

The decisions of the consumers and the merchants at stage 4 and 5 remain unchanged. At stage 3, the best responses of the Issuer and of merchant 1 are the same as in section 4.2. However, perfect competition leads the acquirers to choose a merchant fee that is equal to the marginal cost of the acquisition activity, that is $m^* = a + c_A$. Simulations show that there exists a maximum level for the interchange fee, that we denote by $\overline{a}$, such that both merchants accept payment cards.

Then, we study the condition under which merchant 1 enters the market for payment card transactions at stage 2. Simulations show that, with perfect competition on the acquisition side, the result of Lemma 3 is not verified any more. A higher interchange fee *increases* merchant 1's incentives to issue private cards. In Section 4.4, we proved that the interchange fee has two effects on merchant 1's entry decision. On the one hand, a higher interchange fee reduces the card fee. This toughens the competition with the Issuer, which lowers the benefits of issuing private cards for merchant 1. On the other hand, a higher interchange fee increases the fee that is paid by merchant 2 each time a consumer pays by card. This raises its perceived marginal cost, which in turn lowers its market share, and provides merchant 1 with higher incentives to issue private cards. With a monopoly on the acquisition side, the first effect was the strongest, because the Acquirer could internalise partly the negative effect of a higher merchant fee on merchant 2's market share. This is not the case if the Acquirers are perfectly competitive, since the merchant fee is equal to the marginal cost of acquisition. In this case, the second effect dominates and a higher interchange fee raises merchant 1's incentives to enter the market for payment card transactions.

Let $\underline{a}$ be the minimum level of interchange fee such that merchant 1 issues private cards. If the payment platform wants to deter entry, it has to set $\left(a^P\right)^* = \underline{a}$. If entry is accommodated, since the Acquirers make zero profit, banks' joint profits are equal to the profit of the Issuer, and increase with the level of interchange fee. Hence, if the payment platform accommodates

entry, it chooses the maximum interchange fee compatible with merchant acceptance, that is $\left(a^P\right)^* = \overline{a}$. The results of Proposition 5 and 6 are modified as follows. If $a^B \leq \underline{a}$, entry is blockaded. If $a^B > \underline{a}$ and if $(\Pi_I)^B(\underline{a}) < (\Pi_I)^{PC}(\overline{a})$, the payment platform accommodates entry, whereas if $(\Pi_I)^B(\underline{a}) \geq (\Pi_I)^{PC}(\overline{a})$, the payment platform deters merchant 1 from issuing private cards.

6 Conclusion

Our paper shows that, with monopolies both on the issuing and on the acquisition side, a payment platform may increase its level of interchange fee to deter a merchant from entering the market for payment card transactions. The effect of the competition with the private card is to reduce the card fee and to increase the cost of card acceptance for the merchant that does not issue private cards.

Further research is needed to understand better other forms of entry accommodation that can be designed by the payment platform. For instance, several merchants have started issuing cards with the support of financial institutions that are members of payment card associations. The payment platform could think of other types of contracts that would enable merchants to "opt-in" the payment system, such as cobranding agreements. Or a large retailer, as the merchant Target in the United-States, could decide to become an issuing member of the payment association. Research is also needed to understand the other opt-out strategies of the merchants. For instance, merchants could decide, as for the Aurore Card in France, to build private networks that compete with payment card associations.

7 Appendix

7.1 Appendix A: Proof of Proposition 1

Assume that both merchants accept payment cards at stage 5. A consumer with benefit b_B, and located at x, buys from merchant 1 if and only if

$$q_1 - p_1 - tx \geq q_2 - p_2 - t(1-x).$$

For $(i;j) \in \{1;2\}^2$ and $i \neq j$, we define

$$w_i = \frac{1}{2} + \frac{1}{2t}(q_i - q_j + p_j - p_i).$$

Consumers such that $b_B \geq f^C$ purchase by card, therefore the demand of card payments for merchant i is $D_i^C = (1-f^C)w_i$. The total demand for card payments is $D_T^C = D_1^C + D_2^C = 1-f^C$. Similarly, consumers such that $b_B \leq f^C$ pay cash, hence the demand for cash payments of merchant i is $D_i^{Cash} = f^C w_i$. Each merchant chooses the price that maximises its profit,

$$\pi_i^{C,C} = (1-f^C)w_i(p_i - c - m + b_S) + f^C w_i(p_i - c).$$

Writing the first order condition, we obtain the prices chosen at the equilibrium of the subgame[17]

$$\begin{aligned} p_i &= c + t + \frac{1}{3}(q_i - q_j) + (m - b_S)(1 - f^C), \\ \pi_i &= \frac{(t + \frac{1}{3}(q_i - q_j))^2}{2t}, \end{aligned}$$

for $(i;j) \in \{1;2\}^2$ and $i \neq j$.

Suppose that merchant 1 deviates from this presumed equilibrium, and decides to refuse payment cards. A consumer with benefit b_B wants to use his payment card if and only if $b_B \geq f^C$. A consumer with benefit $b_B \geq f^C$ located at x buys from merchant 1 if and only if :

$$q_1 - p_1 - tx \geq q_2 - p_2 - t(1-x) + b_B - f^C.$$

Aggregating over all customers such that $b_B \geq f^C$, we obtain the demand of the consumers who wish to use their payment cards, and still choose to shop at merchant 1, even if the latter refuses cards:

$$(1 - f_C)w_1 - \frac{1}{4t}(1 - f^C)^2.$$

The demand of the consumers who wish to use cash and choose merchant 1 is equal to $f^C w_1$. Merchant 1 and merchant 2 choose respectively the prices p_1 and p_2 that maximise their profits:

$$\begin{aligned} \pi_1^{C,C} &= \left(w_1 - \frac{1}{4t}(1 - f^C)^2\right)(p_1 - c), \\ \pi_2^{C,C} &= \left((1 - f^C)w_2 + \frac{1}{4t}(1 - f^C)^2\right)(p_2 - c + b_S - m) + f_C w_2(p_2 - c). \end{aligned}$$

Solving for the first order conditions yields[18]:

[17] The second order condition is always satisfied.

$$\begin{aligned} 2p_1 &= t + c + q_1 - q_2 + p_2 - \frac{(1-f^C)^2}{2}, \\ 2p_2 &= t + c + q_2 - q_1 + p_1 + (m - b_S)(1-f^C) + \frac{(1-f^C)^2}{2}. \end{aligned}$$

At the equilibrium, we obtain:

$$\begin{aligned} p_1 &= t + c + \frac{1}{3}(q_1 - q_2 + (m-b_S)(1-f^C) - \frac{(1-f^C)^2}{2}), \\ p_2 &= t + c + \frac{1}{3}(q_2 - q_1 + 2(m-b_S)(1-f^C) + \frac{(1-f^C)^2}{2}), \\ \pi_1^{C,C} &= \frac{1}{2t}\left[t + \frac{1}{3}(q_1 - q_2 + (m-b_S)(1-f^C) - \frac{(1-f^C)^2}{2})\right]^2, \\ \pi_2^{C,C} &= \frac{1}{2t}\left[t + \frac{1}{3}(q_2 - q_1 - (m-b_S)(1-f^C) + \frac{(1-f^C)^2}{2})\right]^2 + \frac{(b_S - m)f^C(1-f^C)^2}{4t}. \end{aligned}$$

Merchant 1 has no incentive to deviate from the equilibrium in which both merchants accept cards if and only if:

$$\frac{1}{2t}(t + \frac{1}{3}(q_1 - q_2))^2 \geq \frac{1}{2t}(t + \frac{1}{3}(q_1 - q_2 + (m-b_S)(1-f^C) - \frac{(1-f^C)^2}{2}))^2,$$

which can be written if $f^C \neq 1$,

$$m \leq b_S + \frac{(1-f^C)}{2}.$$

This condition is the same for merchant 2.

At stage 3, the issuer and the acquirer maximise their profits,

$$\begin{aligned} \Pi_I &= (1-f^C)(f^C + a^P - c_I), \\ \Pi_A &= (1-f^C)(m - a^P - c_A), \end{aligned}$$

with respect to f^C and m, respectively, subject to the constraint,

$$m \leq b_S + \frac{1-f^C}{2}.$$

The constraint is binding for the acquirer since $\frac{d\Pi_A}{dm} = 1 - f^C \geq 0$. Therefore, the best response of the acquirer is to choose

$$m = b_S + \frac{(1-f^C)}{2}.$$

[18] The second order conditions are always satisfied.

Solving for the first-order condition of profit maximisation for the issuer yields the best response[19]

$$f^C = \frac{1 + c_I - a^P}{2}.$$

In this case, the optimal merchant fee is

$$m = b_S + \frac{1 - c_I + a^P}{4}$$

At stage 1, the payment card system chooses the interchange fee that maximises banks' joint profits,

$$\Pi_I + \Pi_A = \frac{1}{2}\left(b_S + \frac{3 + c_I - a^P}{4} - (c_I + c_A)\right)\left(1 + a^P - c_I\right).$$

Solving for the first order condition yields[20]

$$\left(b_S + \frac{3 + c_I - a^P}{4} - (c_I + c_A)\right) - \frac{1}{4}\left(1 - c_I + a^P\right) = 0,$$

and the optimal interchange fee is

$$a^P = 2(b_S - c_A) + 1 - c_I.$$

The optimal transaction fees are then

$$f^C = c_I + c_A - b_S,$$

and

$$m = \frac{3b_S + 1 - c_A - c_I}{2}.$$

7.2 Appendix B: Equilibrium profits

Both merchants accept cards (Section 4.1.1) If both merchants accept cards, the equilibrium profits are

$$\begin{aligned}\pi_1^{C,C} &= \frac{1}{2t}\left(t + \frac{1}{3}\left(\Delta q + \frac{(\Delta f)^2}{2} + \left(f^{PC} - c_M\right)\left(1 - f^{PC}\right) + (\Delta f + m)(1 - f^C) + b_S \Delta f\right)\right)^2 \\ &\quad + \frac{(\Delta f)(f^{PC} + b_S - c_M) f^{PC}}{2t} \times \left(\frac{\Delta f}{2} + 1 - f^C\right),\end{aligned}$$

[19] The second order condition is verified.
[20] The second order condition is verified.

and

$$\pi_2^{C,C} = \frac{1}{2t}\left(t+\frac{1}{3}\left(-\Delta q-\frac{(\Delta f)^2}{2}-\left(f^{PC}-c_M\right)\left(1-f^{PC}\right)-(\Delta f+m)(1-f^C)-b_S\Delta f\right)\right)^2 + \frac{(m-b_S)(1-f^C)(f^C-f^{PC})\left(f^C+f^{PC}\right)}{4t}.$$

Merchant 1 does not accept payment cards, while merchant 2 accepts them (Section 4.1.2) If merchant 1 does not accept cards, and merchant 2 accepts them, the equilibrium profits are

$$\pi_1^{NC,C} = \frac{1}{2t}\left(t+\frac{1}{3}\left(\Delta q+(m-b_S)\left(1-f^C\right)-\frac{(\Delta f)^2}{2}+(f^{PC}+b_S-c_M)(1-f^{PC})+\left(1-f^{PC}\right)\Delta f\right)\right)^2 + \frac{(f^{PC}-c_M+b_S)}{4t}(\Delta f)(1-f^{PC})(f^C+f^{PC}),$$

and

$$\pi_2^{NC,C} = \frac{1}{2t}\left(t+\frac{1}{3}\left(-\Delta q-(m-b_S)\left(1-f^C\right)+\frac{(\Delta f)^2}{2}-(f^{PC}+b_S-c_M)(1-f^{PC})-\left(1-f^{PC}\right)\Delta f\right)\right)^2 \frac{f^C(\Delta f)(m-b_S)}{2t}\left(1-\frac{f^{PC}+f^C}{2}\right).$$

Merchant 1 accepts all cards, while merchant 2 refuses them, or both merchants refuse cards (Sections 4.1.3 and 4.1.4) If merchant 1 accepts all cards and merchant 2 refuses payment cards or if both merchants refuse cards, the equilibrium profits are

$$\pi_1^{C,NC}=\pi_1^{NC,NC}=\frac{1}{2t}\left(t+\frac{1}{3}\left(\Delta q+\frac{(1-f^{PC})^2}{2}+(f^{PC}+b_S-c_M)(1-f^{PC})\right)\right)^2+\frac{f^{PC}(f^{PC}+b_S-c_M)(1-}{4t}$$

and

$$\pi_2^{C,NC}=\pi_2^{NC,NC}=\frac{1}{2t}\left(t+\frac{1}{3}\left(-\Delta q-\frac{(1-f^{PC})^2}{2}-(f^{PC}+b_S-c_M)(1-f^{PC})\right)\right)^2.$$

7.3 Appendix C: Proof of Proposition 2

7.3.1 Appendix C1: π_1 decreases with f^{PC} if $f^{PC} < f^C$.

Both merchants accept cards If $f^{PC} < f^C$, consumers who shop at merchant's 1 pay with the private card, hence, whether merchant 1 accepts cards or not is irrelevant.

Merchant 1's profit is

$$\pi_1^{C,C} = \frac{1}{2t}\left(t+\frac{1}{3}\left(\Delta q+\frac{(\Delta f)^2}{2}+\left(f^{PC}-c_M\right)\left(1-f^{PC}\right)+(m+\Delta f)(1-f^C)+b_S(\Delta f)\right)\right)^2 + \frac{f^{PC}(f^{PC}+b_S-c_M)(\Delta f)}{2t}\left[(1-f^C)+\frac{\Delta f}{2}\right].$$

Derivating with respect to f^{PC}, we obtain

$$\frac{\partial \pi_1^{C,C}}{\partial f^{PC}} = \frac{-H_1}{36t},$$

where

$$H_1 = 4\Delta f\left(b_S\right)^2 + Xb_S + Y,$$

$$X_1 = X_1\left(t, \Delta q, f^C, f^{PC}, m, c_M\right),$$

$$Y_1 = Y_1\left(t, \Delta q, f^C, f^{PC}, m, c_M\right).$$

We want to prove that $H_1 \geq 0$, which would lead that $\partial \pi_1^{C,C}/\partial f^{PC} \leq 0$. We do it in a few steps. First, we prove that

$$\left.\frac{\partial \pi_1}{\partial f^{PC}}\right|_{f^{PC}=0} \leq 0.$$

Indeed, we have

$$\left.\frac{\partial \pi_1^{C,C}}{\partial f^{PC}}\right|_{f^{PC}=0} = -\frac{(b_S-c_M)}{36t}K_1,$$

where $K_1 = 12t + 4\Delta q + 4m\left(1-f^C\right) + 4\left(b_S-c_M\right) - \left[4b_S\left(1-f^C\right)+14f^C-7\left(f^C\right)^2\right]$. Given that $b_S \leq 1$ and $f^C \in [0,1]$, it can be shown that the term into brackets is always strictly lower than 8. Hence, $3t + \Delta q \geq 2$ implies that $K_1 > 0$. Since $b_S > c_M$ by assumption, we have

$$\left.\frac{\partial \pi_1^{C,C}}{\partial f^{PC}}\right|_{f^{PC}=0} \leq 0.$$

Second, we prove that

$$\left.\frac{\partial^2 \pi_1^{C,C}}{\partial \left(f^{PC}\right)^2}\right|_{f^{PC}=0} \leq 0. \tag{8}$$

Indeed, we have

$$\left.\frac{\partial^2 \pi_1^{C,C}}{\partial \left(f^{PC}\right)^2}\right|_{f^{PC}=0} = \frac{-M_1}{9t},$$

where

$$M_1 = 3t+\Delta q+9\left(b_S-c_M\right)+\left(b_S-c_M\right)\left(1-\left(b_S-c_M\right)\right)+m\left(1-f^C\right)-\left[\left(b_S+4f^C\right)\left(1-f^C\right)+4f^C\right].$$

The term into brackets is lower than 5. Hence, if $3t+\Delta q \geq 5$, and given that $b_S \geq c_M$, we have $M_1 \geq 0$, which implies that (8) holds.

Third, we find that the third-order derivative of π_1, denoted by $\pi_1^{(3)}$, has the sign of $114f^{PC}-54+33\left(b_S-c_M\right)$. When $f^{PC}=0$, we have $\pi_1^{(3)}<0$ as $b_S-c_M \leq 1$. When $f^{PC}=1$, we have $\pi_1^{(3)}>0$ as $b_S-c_M>0$. Therefore, $\pi_1^{(3)}<0$ for low values of f^{PC} and $\pi_1^{(3)}>0$ for high values of f^{PC}, which implies that $\pi_1^{(2)}$ is first decreasing then increasing.

Given these properties, we know that either $\pi_1^{(1)}$ is always negative, or it is first negative then positive (as a function of f^{PC}). The second case occurs when $\pi_1^{(2)}$ becomes positive for high values of f^{PC} and $\pi_1^{(1)}$ increases sufficiently to become positive. Therefore, the global optimum of $\pi_1(f^{PC})$ when $f^{PC} \in \left[0, f^C\right)$ is either 0 or f^{C-}. We have

$$\pi_1^{C,C}(0) = \frac{1}{2t}\left[t+\frac{1}{3}\left(\Delta q+m\left(1-f^C\right)+\left(f^C-c_M\right)\left(1-f^C\right)+\left(b_S-c_M\right)f^C+\frac{\left(f^C\right)^2}{2}\right)\right]^2,$$

and

$$\pi_1^{C,C}\left(f^{C-}\right) = \frac{1}{2t}\left[t+\frac{1}{3}\left(\Delta q+m\left(1-f^C\right)+\left(f^C-c_M\right)\left(1-f^C\right)\right)\right]^2,$$

hence $\pi_1^{C,C}(0) > \pi_1^{C,C}\left(f^{C-}\right)$ if and only if $f^C+2\left(b^S-c_M\right)>0$, which is true (for all $f^C \geq 0$) since $b^S > c_M$.

Merchant 2 refuses all payment cards If $f^{PC} \leq f^C$, merchant 1's profit is

$$\pi_1^{C,NC} = \frac{1}{2t}\left(t+\frac{1}{3}\left(\Delta q+\frac{(1-f^{PC})^2}{2}+(f^{PC}+b_S-c_M)(1-f^{PC})\right)\right)^2+\frac{f^{PC}(f^{PC}+b_S-c_M)(1-f^{PC})^2}{4t}.$$

Derivating with respect to f^{PC}, we obtain

$$\frac{\partial \pi_1^{C,NC}}{\partial f^{PC}} = \frac{-H_2}{36t},$$

where

$$\begin{aligned} H_2 &= 4(1-f^{PC})(b_S)^2+X_2b_S+Y_2, \\ X_2 &= X_2\left(t,\Delta q,f^{PC},c_M\right), \\ Y_2 &= Y_2\left(t,\Delta q,f^{PC},c_M\right). \end{aligned}$$

We want to prove that for any $b_S \geq 0$, $\Delta q \geq 0$, $f^{PC} \in [0,1]$, $\partial \pi_1^{C,NC}/\partial f^{PC} \leq 0$. We use the same steps as above. First, we prove that

$$\left.\frac{\partial \pi_1^{C,NC}}{\partial f^{PC}}\right|_{f^{PC}=0} \leq 0.$$

Indeed, we have

$$\left.\frac{\partial \pi_1^{C,NC}}{\partial f^{PC}}\right|_{f^{PC}=0} = -\frac{(b_S - c_M)}{36t} K_2,$$

where $K_2 = 12t + 4\Delta q - 7 - 8c_M + 4(b_S + c_M)$. Since $4\Delta q + 4(b_S + c_M) > 0$, and since $c_M < 1$, $t + \Delta q/3 \geq 5/4$ implies that $K_2 > 0$. Since $b_S > c_M$ by assumption, we have

$$\left.\frac{\partial \pi_1^{C,NC}}{\partial f^{PC}}\right|_{f^{PC}=0} \leq 0.$$

Second, we prove that

$$\left.\frac{\partial^2 \pi_1^{C,NC}}{\partial \left(f^{PC}\right)^2}\right|_{f^{PC}=0} \leq 0. \tag{B1}$$

Indeed, we have

$$\left.\frac{\partial^2 \pi_1^{C,NC}}{\partial \left(f^{PC}\right)^2}\right|_{f^{PC}=0} = \frac{-M_2}{36t},$$

where

$$M_2 = 12t + 4\Delta q - 16 + (b_S - c_M)(40 - 4b_S + 4c_M).$$

Since $b_S < 1$, then $40 - 4b_S + 4c_M > 0$. Hence, given that $b_S > c_M$, by Assumption 1, we have $M_2 \geq 0$, which implies that (B1) holds. Third, we find that the third-order derivative of π_1, denoted by $\pi_1^{(3)}$, has the sign of $114 f^{PC} + 33\,(b_S - c_M) - 54$.

When $f^{PC} = 0$, we have $\pi_1^{(3)} < 0$ as $b_S - c_M \leq 1$. When $f^{PC} = 1$, we have $\pi_1^{(3)} > 0$ as $b_S - c_M \geq 0$. Therefore, $\pi_1^{(3)} < 0$ for low values of f^{PC} and $\pi_1^{(3)} > 0$ for high f^{PC}. It implies that $\pi_1^{(2)}$ is first decreasing then increasing. Given these properties, we know that either $\pi_1^{(1)}$ is always negative, or it is first negative then positive (as a function of f^{PC}). The second case occurs when $\pi_1^{(2)}$ becomes positive for high values of f^{PC} and $\pi_1^{(1)}$ increases sufficiently to become positive. Therefore, the global optimum of $\pi_1^{C,NC}\left(f^{PC}\right)$ when $f^{PC} \in \left[0, f^{C}\right)$ is either 0 or f^{C-}. We have

$$\pi_1^{C,NC}(0) = \frac{1}{2t}\left[t + \frac{1}{3}\left(\Delta q + b_S - c_M + \frac{1}{2}\right)\right]^2,$$

and

$$\pi_1^{C,NC}\left(f^{C-}\right) = \frac{1}{2t}\left[t + \frac{1}{3}\left(\Delta q + b_S - c_M + \frac{1}{2} - f^{C}(b_S - c_M) - \frac{(f^{C})^2}{2}\right)\right]^2.$$

Since $b^S - c_M > 0$, we have $\pi_1(0) > \pi_1\left(f^{C-}\right)$.

To sum up, in cases 1-4, the global maximum of $\pi_1^{C,NC}\left(f^{PC}\right)$ over $\left[0, f^C\right)$ is obtained at $f^{PC} = 0$.

7.3.2 Appendix C2: Merchant 1 undercuts f^C by setting $f^{PC} < f^C$

We show that, in all cases, merchant 1 always makes more profit if he undercuts the Issuer by setting $f^{PC} < f^C$.

Case 1: Both merchants accept payment cards. If $f^{PC} > f^C$, merchant 1 makes profit

$$\pi_1^{C,C} = \frac{1}{2t}\left(t + \frac{\Delta q}{3}\right)^2.$$

If $f^{PC} < f^C$, we know from Appendix B1 that merchant 1's profit is maximum for $f^{PC} = 0$, in which case he makes

$$\pi_1^{C,C} = \frac{1}{2t}\left(t + \frac{1}{3}\left(\Delta q + \frac{(f^C)^2}{2} + (b_S - c_M) + (f^C + m - b_S)(1 - f^C)\right)\right)^2.$$

Let $\gamma = m - a - c_A$ and $\delta = f^C + a - c_I$ be the Acquirer's and Issuer's margins, respectively. Since the margins are positive, we have $\gamma \geq 0$ and $\delta \geq 0$. We also have $f^C + m - b_S = \delta + \gamma + c_I + c_A - b_S$. Since $b_S \leq c_I + c_A$ by assumption, it follows that $f^C + m - b_S \geq 0$. As we have $b_S > c_M$ too, then merchant 1 makes more profit if he undercuts the Issuer by setting $f^{PC} < f^C$.

Case 2: Merchant 2 is the only one who accepts cards. If $f^{PC} < f^C$, merchant 1 makes profit

$$\begin{aligned}\pi_1^{NC,C} &= \frac{1}{2t}\left(t + \frac{1}{3}\left(\Delta q + \frac{(\Delta f)^2}{2} + (f^{PC} - c_M)(1 - f^{PC}) + (\Delta f + m)(1 - f^C) + b_S \Delta f\right)\right)^2 \\ &\quad + \frac{(\Delta f)(f^{PC} + b_S - c_M) f^{PC}}{2t} \times \left(\frac{\Delta f}{2} + 1 - f^C\right),\end{aligned}$$

whereas if $f^{PC} > f^C$, merchant 1 makes profit

$$\begin{aligned}\pi_1^{NC,C} &= \frac{1}{2t}\left(t + \frac{1}{3}\left(\Delta q + \frac{(\Delta f)^2}{2} + (f^{PC} - c_M)(1 - f^{PC}) + (\Delta f + m)(1 - f^C) + b_S \Delta f\right)\right)^2 + \\ &\quad \frac{(f^{PC} + b_S - c_M)}{4t}(\Delta f)(1 - f^{PC})(f^C + f^{PC}).\end{aligned}$$

We show that merchant 1 makes more profit if $f^{PC} < f^C$. Since $b_S \geq c_M$, we have $f^{PC} + b_S - c_M \geq 0$. If $f^{PC} < f^C$, then $(\Delta f) \geq 0$. So,

$$\frac{(\Delta f)(f^{PC} + b_S - c_M)f^{PC}}{2t} \times \left(\frac{\Delta f}{2} + 1 - f^C\right) \geq 0.$$

If $f^{PC} > f^C$, then $(\Delta f) \leq 0$. So, we have

$$\frac{(f^{PC} + b_S - c_M)}{4t}(\Delta f)(1 - f^{PC})(f^C + f^{PC}) \leq 0.$$

Therefore, merchant 1 makes more profit if he chooses $f^{PC} < f^C$.

Case 3: Merchant 1 is the only one who accepts cards. If $f^{PC} < f^C$, merchant 1 makes profit

$$\pi_1^{C,NC} = \frac{1}{2t}\left(t + \frac{1}{3}\left(\Delta q + \frac{(1 - f^{PC})^2}{2} + (f^{PC} + b_S - c_M)(1 - f^{PC})\right)\right)^2 + \frac{f^{PC}(f^{PC} + b_S - c_M)(1 - f^{PC})^2}{4t},$$

whereas if $f^{PC} > f^C$, he makes profit

$$\pi_1^{C,NC} = \frac{1}{2t}\left[t + \frac{1}{3}(\Delta q + (b_S - m)(1 - f^C) + \frac{(1 - f^C)^2}{2})\right]^2 + \frac{(b_S - m)f^C(1 - f^C)^2}{4t}.$$

Notice that this situation is possible if and only if the non deviation condition in the Benchmark Case is not verified, that is, if we have

$$m \geq b_S + \frac{1 - f^C}{2}.$$

Therefore, in case 3, if $f^{PC} > f^C$, we have $(b_S - m) \leq 0$. So, $\frac{(b_S - m)f^C(1 - f^C)^2}{4t} \leq 0$. Consequently, to prove that merchant 1 makes more profit if he undercuts f^C, it suffices to prove that

$$C \geq D,$$

where

$$C = \frac{(1 - f^{PC})^2}{2} + (f^{PC} + b_S - c_M)(1 - f^{PC}),$$

and

$$D = (b_S - m)(1 - f^C) + \frac{(1 - f^C)^2}{2}.$$

Rearranging C, and using that $1 - f^{PC} = 1 - f^C + \Delta f$, we obtain

$$C = \frac{(1-f^C)^2}{2} + (f^{PC} + b_S - c_M)(1-f^C) + \frac{(\Delta f)^2}{2} + (f^{PC} + b_S - c_M + 1 - f^C)(\Delta f).$$

Since $(f^{PC} + b_S - c_M)(1-f^C) \geq (b_S - m)(1-f^C)$, and since $\frac{(\Delta f)^2}{2} + (f^{PC} + b_S - c_M + 1 - f^C)(\Delta f) \geq 0$, we have $C \geq D$. Therefore, merchant 1 makes more profit if he chooses $f^{PC} < f^C$.

Case 4: Both merchants refuse payment cards. This case is not relevant, as both merchants refuse cards (and hence, merchant 1's profit does not depend on whether $f^{PC} < f^C$ or $f^{PC} > f^C$).

To sum up, in all cases, merchant 1 makes more profit if he undercuts f^C by setting $f^{PC} < f^C$.

7.4 Appendix D: Proof of Lemma 1

Assume that merchant 1 sets $f^{PC} = 0$. Merchant 2 does not change his decision to accept cards if and only if his profit is higher if it accepts cards than if it does not, that is,

$$\frac{1}{2t}\left(t + \frac{1}{3}\left(-\Delta q - \frac{(f^C)^2}{2} + c_M - (f^C + m)(1-f^C) - b_S f^C\right)\right)^2 + \frac{(m-b_S)(1-f^C)(f^C)^2}{4t} \geq \frac{1}{2t}\left(t + \frac{1}{3}\left(-\Delta q - \frac{1}{2} - (b_S - c_M)\right)\right)^2.$$

This condition is equivalent to $g(m, f^C) \geq 0$, where

$$\begin{aligned} g(m, f^C) \quad = \quad & \left(t + \frac{1}{3}\left(-\Delta q - \frac{(f^C)^2}{2} + c_M - (f^C + m)(1-f^C) - b_S f^C\right)\right)^2 \\ & + \frac{(m-b_S)(1-f^C)(f^C)^2}{2} \\ & - \left(t + \frac{1}{3}\left(-\Delta q - \frac{1}{2} - (b_S - c_M)\right)\right)^2. \end{aligned}$$

Condition under which the demand is positive We know that, if both merchants accept cards and $f^{PC} = 0$, the demand for card payments at merchant 2's is

$$\begin{aligned} D_2^C &= (1-f^C)w_2 - \frac{1}{2t}(1-f^C)f^C \\ &= \frac{1}{2t}(1-f^C)\left(t + \frac{1}{3}(-\Delta q - (b_S+1+f^C)f^C + c_M - m(1-f^C))\right). \end{aligned} \tag{C1}$$

Therefore, we have $D_2^C \geq 0$ if and only if

$$m \leq \overline{m},$$

where

$$\overline{m} = \frac{3t - \Delta q + c_M - (b_S + 1 + f^C)f^C}{(1-f^C)}.$$

Existence and characterization of $\widetilde{m}(f^C)$ First, we show that there exists an $\widetilde{m}(f^C)$ such that merchant 2 does not deviate from the equilibrium in which he accepts cards for $m \leq \widetilde{m}(f^C)$. Indeed, note that g is a convex polynomial function of m of degree 2, as $\partial^2 g/\partial m^2 = 2\left(1-f^C\right)^2/9 > 0$. Besides, we have

$$g(\overline{m}, f^C) = \frac{\left(f^C\right)^4}{4} + \frac{(f^C)^2}{2}\left[3t - \Delta q + c_M - b_S - f^C(1+f^C)\right] - \left(\frac{3t - \Delta q - b_S + c_M - 1/2}{3}\right)^2, \tag{Cx}$$

and $g(\overline{m}, f^C) \leq 0$ for sufficiently high t.

Indeed, for sufficiently high t, we prove that $g(\overline{m}, f^C)$ is increasing in f^C. If $g(\overline{m}, 1) \leq 0$, we will have shown that $g(\overline{m}, f^C) \leq 0$ for sufficiently high t. We have

$$\frac{\partial g(\overline{m}, f^C)}{\partial f^C} = f^C\left[3t - \Delta q - b_S + c_M - \left(f^C\right)^2 - \frac{3}{2}f^C\right].$$

Since $c_M - b_S \geq -1$, $\left(f^C\right)^2 + \frac{3}{2}f^C \in [0, 5/2]$, to have $\frac{\partial g(\overline{m}, f^C)}{\partial f^C} \geq 0$, it suffices that

$$t \geq \frac{\Delta q}{3} + \frac{7}{6},$$

which is true by Assumption 1. Replacing for $f^C = 1$ in (Cx), we obtain

$$g(\overline{m}, 1) = \frac{-1}{18}\left(6t - 2\Delta q - 7 + 2c_M - 2b_S\right)\left(3t - \Delta q - 2 - b_S + c_M\right).$$

To have $g(\overline{m}, 1) \leq 0$, it is sufficient that both parenthesis are positive, which is the case by Assumption 1.

Hence, $g(\overline{m}, f^C) \leq 0$ for sufficiently high t.

Then, note that $b_S + 1 - f^C \leq \overline{m}$ for sufficiently high t. Indeed, this condition is equivalent to

$$3t - \Delta q - b_S + c_M - f^C(1 + f^C) - (1 - f^C)^2 \geq 0,$$

which is true if $t \geq (\Delta q + 4)/3$, by Assumption 1.

We have

$$g(b_S + \frac{3(1-f^C)}{4}, f^C) = \frac{-(1-f^C)^2}{144}\left(24t - 8\Delta q - 55\left(f^C\right)^2 - 8\left(b_S - c_M\right) + 5 - 2f^C\right). \quad \text{(C1)}$$

We have $\left(f^C\right)^2 \leq 1$, $(b_S - c_M) \in [0, 1]$ and $2f^C \leq 2$, therefore, (C1) is negative if $t > \frac{\Delta q}{3} + \frac{5}{2}$, which is true by Assumption 1.

Finally, we obtain that

$$g(b_S + \frac{1-f^C}{2}, f^C) = \frac{(1-f^C)^2(f^C)^2}{4} \geq 0.$$

This shows that $g\left(m, f^C\right)$ is first positive then negative over $[0, \overline{m}]$, and that it crosses $y = 0$ only once, at $\widetilde{m}(f^C)$. Besides, since $g(b_S + \frac{3(1-f^C)}{4}, f^C) < 0$ and $g(b_S + \frac{1-f^C}{2}, f^C) \geq 0$, we have

$$\widetilde{m}(f^C) \in \left(b_S + \frac{(1-f^C)}{2}; b_S + \frac{3(1-f^C)}{4}\right).$$

7.5 Appendix E: Proof of Proposition 3

The first order conditions of profit maximisation for the Acquirer and the Issuer are

$$\frac{dD_2^C}{dm}(m - a^P - c_A) + D_2^C = 0, \quad \text{(E1)}$$

and

$$\frac{dD_2^C}{df^C}(f^C + a^P - c_I) + D_2^C = 0, \quad \text{(E2)}$$

respectively. Proposition 2 shows that $f^{PC} = 0$ is a dominant strategy for merchant 1. Therefore, we replace for $f^{PC} = 0$ in (E1) and (E2). We have

$$D_2^C = \frac{1}{2t}(1 - f^C)\left(t + \frac{1}{3}(-\Delta q - (b_S + 1 + f^C)f^C + c_M - m(1 - f^C))\right).$$

We define

$$R = \frac{2t}{(1-f^C)}D_2^C. \quad \text{(E3)}$$

Since $\frac{dD_2^C}{dm} = \frac{-(1-f^C)^2}{6t}$ and $\frac{dD_2^C}{df^C} = \frac{-R}{2t} + \frac{(1-f^C)}{6t} \times (m - 1 - b_S - 2f^C)$, by simplifying (E1) and (E2), we obtain

$$\begin{aligned} \frac{-(1-f^C)}{3}(m - a^P - c_A) + R &= 0, \\ (f^C + a^P - c_I)(-R + \frac{(1-f^C)}{3} \times (m - b_S - 1 - 2f^C)) + (1-f^C)R &= 0. \end{aligned}$$

Before solving for the equilibrium, we start by showing that the Issuer's and the Acquirer's profit functions are concave.

7.5.1 Appendix E1: Concavity of profit functions

Writing the second derivative of Π_A with respect to m, we obtain

$$\frac{\partial^2 \Pi_A}{\partial^2 m} = \frac{-(1-f^C)^2}{3t} < 0,$$

so the second order condition for the Acquirer is verified.

Writing the second derivative of Π_I with respect to f^C, we obtain

$$\begin{aligned} \frac{\partial^2 \Pi_I}{\partial^2 f^C} &= -1 - \frac{1}{3t}\left[(f^C + a^P - c_I)(m - b_S - 3f^C) - \Delta q - (b_S + 1 + f^C)f^C\right] \\ &\quad + \frac{1}{3t}\left[-c_M + (1-f^C)(2m - b_S - 1 - 2f^C)\right]. \end{aligned}$$

The third derivative of Π_I with respect to f^C is given by

$$\frac{\partial^3 \Pi_I}{\partial^3 f^C}(m, f^C) = \frac{1}{t}\left(4f^C + a^P - c_I - m + b_S\right).$$

Replacing for $f^C = 0$ yields

$$\frac{\partial^3 \Pi_I}{\partial^3 f^C}(m, 0) = \frac{1}{t}\left(a^P - c_I - m + b_S\right).$$

Since the Acquirer's profit must be positive, we have $m - a^P - c_A \geq 0$. So

$$a^P - c_I - m + b_S \leq b_S - c_I - c_A.$$

Since $b_S - c_I - c_A < 0$ by assumption, we have

$$\frac{\partial^3 \Pi_I}{\partial^3 f^C}(m, 0) < 0.$$

Replacing for $f^C = 1$ yields

$$\frac{\partial^3 \Pi_I}{\partial^3 f^C}(m,1) = \frac{1}{t}\left(4 + a^P - c_I - m + b_S\right) = \frac{1}{t}(3 + a^P - c_I + 1 + b_S - m).$$

In Appendix C, we proved that $\widetilde{m}(f^C) \leq b_S + 3\left(1 - f^C\right)/4$, so $m \leq \widetilde{m}(f^C) \leq 1 + b_S$. Since the margin of the Issuer, $f^C + a^P - c_I$, must be positive, and $f^C \in [0,1]$, we have $3 + a^P - c_I \geq 0$. So

$$\frac{\partial^3 \Pi_I}{\partial^3 f^C}(m,1) > 0.$$

Therefore, as $\frac{\partial^3 \Pi_I}{\partial^3 f^C}$ is increasing with f^C, there exists a unique $\widetilde{f}^C \in (0;1)$ such that $\frac{\partial^3 \Pi_I}{\partial^3 f^C}(m, f^C) > 0$ if $f^C > \widetilde{f}^C$ and $\frac{\partial^3 \Pi_I}{\partial^3 f^C}(m, f^C) \leq 0$ otherwise. To show that $\frac{\partial^2 \Pi_I}{\partial^2 f^C}(m, f^C) \leq 0$, it suffices to prove that $\frac{\partial^2 \Pi_I}{\partial^2 f^C}(m,0) \leq 0$, and that $\frac{\partial^2 \Pi_I}{\partial^2 f^C}(m,1) \leq 0$. Replacing for $f^C = 0$ yields

$$\frac{\partial^2 \Pi_I}{\partial^2 f^C}(m,0) = \frac{-1}{3t}(3t - \Delta q + (m - b_S)(a^P - c_I) + 1 + b_S - m + c_M - m).$$

We know from Appendix C that $1 + b_S - m > 0$. We now show that $\left|a^P - c_I\right| \leq 1$. Since the Issuer's margin is positive, $1 + a^P - c_I \geq f^C + a^P - c_I \geq 0$. So $-1 \leq a^P - c_I$. Since the Acquirer's margin is positive, we have $a^P - c_I \leq m - c_A - c_I$. Hence, $a^P - c_I \leq 1 + b_S - c_A - c_I$. By assumption, $b_S - c_A - c_I \leq 0$. Therefore, $-1 \leq a^P - c_I \leq 1$. Since $\left|a^P - c_I\right| \leq 1$ and $|m - b_S| \leq 1$ then $(m - b_S)(a^P - c_I) \geq -1$. We also know that $-m \geq -b_S - 1 \geq -2$.

Therefore, to prove that $\partial^2 \Pi_I / \partial^2 f^C(m, 0, f^{PC}) \leq 0$, it suffices that $3t - \Delta q - 3 \geq 0$, which is equivalent to $t \geq \Delta q/3 + 1$. This is true by Assumption 1. Replacing for $f^C = 1$ yields

$$\frac{\partial^2 \Pi_I}{\partial^2 f^C}(m,1) = \frac{-1}{3t}(3t - \Delta q + m(1 + a^P - c_I) - (3 + b_S)(1 + a^P - c_I) - (b_S - c_M) - 2)).$$

Since $3 + b_S \leq 4$, and $1 + a^P - c_I \leq 2$, we have $-(3 + b_S)(1 + a^P - c_I) \geq -8$. We also have $-(b_S - c_M) \geq -1$. Therefore, to show that $\frac{\partial^2 \Pi_I}{\partial^2 f^C}(m,1) \leq 0$, it suffices that $3t - \Delta q - 11 \geq 0$, which is equivalent to $t \geq \frac{\Delta q}{3} + \frac{11}{3}$. This is true by Assumption 1.

To sum up, by Assumption 1, Π_I and Π_A are concave with respect to f^C and m, respectively.

7.5.2 Appendix E2: The best response of the Issuer is strictly positive.

We have that

$$\left.\frac{\partial \Pi_I}{\partial f^C}\right|_{f^C=0} = \frac{1}{2t}\left[\left(1 + a^P - c_I\right)\left(t + \frac{1}{3}(-\Delta q + c_M - b_S)\right) + \frac{(b_S + 1 - m)}{3}\right].$$

Since $\widetilde{m}(f^C) \leq b_S + \frac{3}{4}$, then $b_S + 1 - m > 0$. Since the margin of the Issuer must be positive, we also know that $1 + a^P - c_I \geq 0$. Since, by Assumption 1, t is sufficiently high such that $t + (-\Delta q + c_M - b_S)/3 \geq 0$, we can conclude that

$$\left.\frac{\partial \Pi_I}{\partial f^C}\right|_{f^C=0} > 0.$$

Therefore, the best response of the Issuer is strictly positive.

7.5.3 Appendix E3: The Acquirer chooses the maximum merchant fee compatible with merchant acceptance.

Assume that the constraint $m \leq \widetilde{m}(f^C)$ is not binding. The best response of the Acquirer is to play m^{BR} which satisfies to the first order condition, that is

$$\frac{-(1-f^C)}{3}(m^{BR} - a^P - c_A) + R = 0,$$

where R is defined in (E3). Rearranging the first order condition, we get

$$\frac{-2(1-f^C)m^{BR}}{3} + t + \frac{1}{3}(-\Delta q - (b_S + 1 + f^C)f^C + c_M + (1-f^C)(a^P + c_A)) = 0.$$

So,

$$m^{BR}(f^C) = \frac{a^P + c_A}{2} + y(f^C),$$

where

$$y(f^C) = \frac{3}{2(1-f^C)}\left(t + \frac{1}{3}(-\Delta q - (b_S + 1 + f^C)f^C + c_M\right).$$

To show that the constraint is binding if the Acquirer plays its best response, it is sufficient to prove that for $m = y(f^C)$, the non deviation condition is not verified, that is, $y(f^C) > \widetilde{m}(f^C)$ (since $m^{BR}(f^C) > y(f^C)$). A simple way of showing that the non deviation condition is violated for $m = y(f^C)$ is to prove that $y(f^C) > b_S + 1 - f^C$, as we know that $b_S + 1 - f^C \geq \widetilde{m}$. We have

$$y - b_S = \frac{3}{2(1-f^C)}\left(t + \frac{1}{3}(-\Delta q - (b_S - c_M) - (1 + f^C)f^C - b_S(1 - f_C)\right).$$

To show that $y(f^C)-b_S > 1-f^C$, it is equivalent to prove that $U \equiv (y-b_S)(1-f^C)-(1-f^C)^2 > 0$. We have

$$U = \frac{3}{2}t - \frac{1}{2}\left(\Delta q + b_S(1-f^C) + (b_S - c_M) + 2(f^C)^2 + (1-f^C)(2-f^C)\right).$$

Since $b_S(1-f^C) < 1$, $(b_S - c_M) < 1$, $2(f^C)^2 < 2$, and $(1-f^C)(2-f^C) < 2$, we have

$$U > \frac{3}{2}t - \frac{\Delta q}{2} - 3.$$

So to have $U > 0$, it suffices that $t > \frac{\Delta q}{3} + 2$. So if Assumption 1 holds, the Acquirer chooses the maximum merchant fee compatible with merchant acceptance.

7.5.4 Appendix E4: The equilibrium

We can now solve for the equilibrium. We start by showing that two lemmas.

Lemma 4 *$\widetilde{m}(f^C)$ is decreasing with f^C.*

Proof. The function $\widetilde{m}(f^C)$ is defined implicitly by the non deviation condition. Using the implicit function theorem, we obtain

$$\frac{\partial \widetilde{m}(f^C)}{\partial f^C} = -\left(\left.\frac{\partial g}{\partial m}\right|_{m=\widetilde{m}}\right)^{-1} \times \left.\frac{\partial g}{\partial f^C}\right|_{m=\widetilde{m}}.$$

Since g is decreasing with m over $[0, \widetilde{m}]$, the sign of $\frac{\partial \widetilde{m}(f^C)}{\partial f^C}$ has the same as $\left.\frac{\partial g}{\partial f^C}\right|_{m=\widetilde{m}}$. Taking the derivative of g with respect to f^C, we obtain

$$\left.\frac{\partial g}{\partial f^C}\right|_{m=\widetilde{m}} = \frac{2Y}{3}(\widetilde{m}(f^C) - (b_S + 1 - f^C)) + \frac{\widetilde{m}(f^C) - b_S}{2}\left(-3(f^C)^2 + 2f^C\right),$$

where

$$Y = t - \frac{1}{3}\left[\Delta q + \frac{1}{2} + (b_S - c_M) - (1-f^C)(\frac{1-f^C}{2} + b_S - \widetilde{m}(f^C))\right].$$

We now show that $\left.\frac{\partial g}{\partial f^C}\right|_{m=\widetilde{m}} < 0$. First, we have $Y \geq 0$ by Assumption 1.

Indeed, since $\widetilde{m}(f^C) - \frac{3}{4}(1-f^C) < 0$, we have $\frac{(1-f^C)}{2} - \widetilde{m}(f^C) > \frac{-(1-f^C)}{4}$. Hence, $\frac{(1-f^C)}{3}\left(\frac{(1-f^C)}{2} - \widetilde{m}(f^C)\right) > \frac{-(1-f^C)^2}{12} \geq \frac{-1}{12}$. Since $\frac{1}{2} + (b_S - c_M) < \frac{3}{2}$, if $t - \frac{1}{3}\left(\Delta q + \frac{3}{2}\right) - \frac{1}{12} \geq 0$, then $Y \geq 0$. Therefore, it suffices that $t \geq \frac{\Delta q}{3} + \frac{7}{12}$, which is true by Assumption 1.

Besides, we have $\widetilde{m}(f^C) - (b_S + 1 - f^C) \leq 0$, so $\left.\frac{\partial g}{\partial f^C}\right|_{m=\widetilde{m}} < 0$ if and only if

$$Y \geq \frac{3}{4} \frac{\widetilde{m}(f^C) - b_S}{1 - f^C + b_S - \widetilde{m}(f^C)} \left(-3(f^C)^2 + 2f^C\right). \tag{F1}$$

We have $1 - f^C + b_S - \widetilde{m}(f^C) \leq b_S - \widetilde{m}(f^C)$ as $\widetilde{m}(f^C) - b_S \geq \left(1 - f^C\right)/2$. Therefore, a sufficient condition for (F1) to hold is

$$Y \geq \frac{3}{4} \left(-3f^C + 2\right) f^C. \tag{F2}$$

We have $\left(-3f^C + 2\right) f^C \leq 1/3$, so (F2) is equivalent to $Y \geq 1/4$, that is,

$$t \geq \frac{\Delta q}{3} + \frac{5}{6},$$

which is true by Assumption 1.

If this condition holds, then $\widetilde{m}(f^C)$ is decreasing with f^C. ■

Lemma 5 *$(f^C)^{BR}$ is increasing with m.*

Proof. The function $(f^C)^{BR}$ is defined implicitly by the first order condition of the maximisation of the Issuer's profit. Using the implicit function theorem, we obtain

$$\frac{\partial (f^C)^{BR}}{\partial m} = - \left(\left.\frac{\partial^2 \Pi_I}{\partial^2 f^C}\right|_{f^C = (f^C)^{BR}} \right)^{-1} \times \left.\frac{\partial^2 \Pi_I}{\partial f^C \partial m}\right|_{f^C = (f^C)^{BR}}.$$

Since we have shown that the second order condition is verified, the sign of $\frac{\partial (f^C)^{BR}}{\partial m}$ is the same as the sign of $\left.\frac{\partial^2 \Pi_I}{\partial f^C \partial m}\right|_{f^C = (f^C)^{BR}}$. Taking the derivative of the first order condition with respect to m, we obtain

$$\left.\frac{\partial^2 \Pi_I}{\partial f^C \partial m}\right|_{f^C = (f^C)^{BR}} = \frac{(1 - (f^C)^{BR})}{6t} \left[3(f^C)^{BR} + 2(a^P - c_I) - 1\right].$$

So, if $(f^C)^{BR} \leq \frac{1 + 2c_I - 2a^P}{3}$, then $\frac{\partial (f^C)^{BR}}{\partial m} \leq 0$, and $\frac{\partial (f^C)^{BR}}{\partial m} > 0$ otherwise. We are going to show that $(f^C)^{BR} > \frac{1 + 2c_I - 2a^P}{3}$, which will prove that $\frac{\partial (f^C)^{BR}}{\partial m} > 0$. To do so, we replace for $f^C = \frac{1 + 2c_I - 2a^P}{3}$ in the first order condition. Since π_I is concave, if

$$\frac{\partial \Pi_I}{\partial f^C}(m, \frac{1 + 2c_I - 2a^P}{3}) > 0$$

then we know that $(f^C)^{BR} > \frac{1 + 2c_I - 2a^P}{3}$.

We have

$$\frac{\partial \Pi_I}{\partial f^C}(m, \frac{1+2c_I-2a^P}{3}) = \frac{\left(1-c_I+a^P\right)H}{162t}, \tag{F3}$$

where

$$H = 27t - 9\Delta q - 9\left(b_S - c_M\right) - 14 + 8a\left(1-c_I\right) + 4\left(a^2 + \left(c_I\right)^2\right) - 8c_I.$$

We have $1 - c_I + a^P \geq f^C - c_I + a^P \geq 0$, therefore, (F3) is positive if and only if $H \geq 0$. Since $0 \leq 9\left(b_S - c_M\right) \leq 9$, and $8c_I \leq 8$, then a sufficient condition for $H \geq 0$ is

$$t \geq \frac{\Delta q}{3} + \frac{31}{27},$$

which is true by Assumption 1. ■

Define $\widetilde{f}$ such that $\widetilde{m}(\widetilde{f}) = 0$. Then we want to prove that $\widetilde{f} > f^*(m=0)$. The card fee $\widetilde{f}$ is defined by the non deviation condition, in which $\widetilde{m}(\widetilde{f}) = 0$, that is

$$\begin{aligned}\left(t + \frac{1}{3}\left(-\Delta q - \frac{1}{2} - b_S + c_M\right)\right)^2 &= \left(t + \frac{1}{3}\left(-\Delta q - \frac{1}{2} - b_S + c_M + b_S(1-\widetilde{f}) - \widetilde{f}(1-\widetilde{f}) - \frac{(\widetilde{f})^2}{2}\right)\right)^2 \\ &\quad - \frac{b_S(1-\widetilde{f})(\widetilde{f})^2}{2}.\end{aligned}$$

This equation can be rewritten as

$$(1-\widetilde{f})\left\{\left[\frac{2}{3}Z + \frac{1}{9} + \left(\frac{1-\widetilde{f}}{9}\right)\left(\frac{1-\widetilde{f}}{2} + b_S\right)\right]\left(\frac{1-\widetilde{f}}{2} + b_S\right) - b_S\frac{\left(\widetilde{f}\right)^2}{4}\right\} = 0,$$

where $Z = t + \frac{1}{3}\left(-\Delta q - \frac{1}{2} - b_S + c_M\right)$. It can be shown that the terms in the second parenthesis are strictly positive. Hence, the only solution of this equation is obtained for $\widetilde{f} = 1$. Therefore, $\widetilde{f} \geq f^*(m=0)$.

Therefore, we have shown that there is a unique equilibrium such that: $\left(f^{PC}\right)^* = 0$; $\left(f^C\right)^* \in (0,1)$; $m^* = \widetilde{m}$.

7.6 Appendix F: Proof of Proposition 4

We already proved that

$$\widetilde{m}(f^C) > b_S + \frac{1-f^C}{2},$$

which shows that for a given f^C, the Acquirer's best response is to choose a higher merchant fee than in the benchmark case.

We now compare $(f^C)^{BR}$ with the best response of the Issuer in the benchmark case, that

is $f^C = \dfrac{1 + c_I - a^P}{2}$. We have

$$\frac{\partial \Pi_I}{\partial f^C}(m, \frac{1 + c_I - a^P}{2}) = \frac{\left(1 - c_I + a^P\right)^2}{24t}\left(m - b_S - 1 - (1 + c_I - a^P)\right).$$

Since $m - b_S - 1 < 0$, and $1 + c_I - a^P > 0$, we have

$$\frac{\partial \Pi_I}{\partial f^C}(m, \frac{1 + c_I - a^P}{2}) < 0.$$

Since π_I is concave, this proves that $(f^C)^{BR} < \dfrac{1 + c_I - a^P}{2}$. So, for a given m, the Issuer chooses a lower transaction fee than in the benchmark case.

To sum up, for a given f^C, the Acquirer's best response is to choose a higher m than in the benchmark case. Besides, for a given m, the Issuer's best response is to set a lower f^C than in the benchmark case, so the equilibrium merchant fee is higher than in the benchmark case, while the card fee is lower.

7.7 Appendix G: Proof of Lemma 2

We start by showing that $(f^C)^{BR}$ is increasing with a^P. The function $(f^C)^{BR}(m, a^P)$ is defined implicitly by the first order condition of the maximisation of the Issuer's profit. Using the implicit function theorem, we obtain

$$\frac{\partial (f^C)^{BR}(m, a^P)}{\partial a^P} = -\left(\frac{\partial^2 \Pi_I}{\partial^2 f^C}(m, f^C, a^P)\right)^{-1} \frac{\partial^2 \Pi_I}{\partial f^C \partial a^P}(m, (f^C)^{BR}, a^P).$$

Since we have shown in Appendix D1 that π_I is concave, the sign of $\dfrac{\partial (f^C)^{BR}(m, a^P)}{\partial a^P}$ is the same as the sign of $\dfrac{\partial^2 \Pi_I}{\partial f^C \partial a^P}(m, (f^C)^{BR}, a^P)$. Taking the derivative of the first order condition with respect to a^P, we obtain

$$\frac{\partial^2 \Pi_I}{\partial f^C \partial a^P}(m, (f^C)^{BR}, a^P) = \frac{1}{2t}\left[-R + \frac{(1 - (f^C)^{BR})}{3}(m - (b_S + 1 + 2(f^C)^{BR}))\right],$$

where R is given by (E3). Since $m < 1 - f^C + b_S$, we know that $m - (b_S + 1 + 2(f^C)^{BR}) < 0$. Since $R \geq 0$, then it follows that $-R + \dfrac{(1 - (f^C)^{BR})}{3}(m - (b_S + 1 + 2(f^C)^{BR})) < 0$. This shows that

$$\frac{\partial^2 \Pi_I}{\partial f^C \partial a^P}(m, (f^C)^{BR}, a^P) < 0.$$

This proves that $(f^C)^{BR}$ is decreasing with a^P. We also know that $(f^{PC})^{BR}(f^C, m)$ does not depend on a^P as it is equal to 0. Besides, $\widetilde{m}(f^C)$ does not depend on a^P either, as the non

deviation condition does not depend on a^P (see the expression of g in Appendix C). So, if a^P increases, the best response of the Acquirer remains unchanged, while the best response of the Issuer decreases. As shown in Lemma 4, $\widetilde{m}(f^C)$ is decreasing with f^C, which proves that $(f^C)^*$ is lower and that m^* is higher if the interchange fee is higher.

7.8 Appendix H: Entry condition

7.8.1 Entry condition

Merchant 1 enters the market if and only if he makes higher profit with the private card at the equilibrium of stage 3 than in the benchmark case. This condition is obtained by replacing for $f^{PC} = 0$, $(f^C)^*$ and m^* in π_1 (case L-1), that is

$$\pi_1^{C,C}(m^*, (f^C)^*, 0) - F \geq \frac{1}{2t}\left(t + \frac{\Delta q}{3}\right)^2,$$

which is equivalent to

$$\left(t + \frac{1}{3}\left(\Delta q - c_M + \left((f^C)^* + m^*\right)\left(1 - (f^C)^*\right) + \frac{((f^C)^*)^2}{2} + b_S(f^C)^*\right)\right)^2 - 2tF \geq \left(t + \frac{\Delta q}{3}\right)^2.$$

7.9 Appendix I: Proof of Lemma 3

Solving for $\widetilde{m}(f^C)$ (see Appendix C), we find that

$$\widetilde{m}(f^C) = b_S + \frac{1 - f^C}{2} - U(f^C),$$

where

$$U(f^C) = \frac{-(2Q + \frac{3}{2}(f^C)^2) + \sqrt{D}}{2(1 - f^C)/3},$$

$$D = (2Q - \frac{3}{2}(f^C)^2) - (1 - f^C)^2(f^C)^2,$$

and

$$Q = t + \frac{1}{3}\left(-\Delta q - \frac{1}{2} - b_S + c_M\right).$$

Since $\widetilde{m}(f^C) \geq b_S + \left(1 - f^C\right)/2$, we have $U(f^C) \leq 0$.

We have $m^* = \widetilde{m}(\left(f^C\right)^*)$ and $\partial\widetilde{m}/\partial a^P = 0$, therefore,

$$\frac{dm^*}{da^P} = \left.\frac{d\widetilde{m}(f^C)}{df^C}\right|_{f^C=(f^C)^*} \times \frac{d\left(f^C\right)^*}{da^P}.$$

Replacing for this expression in $(EC)'(a^P)$ and replacing for $\widetilde{m}(\left(f^C\right)^*)$, we find that

$$(EC)'(a^P) = \frac{2}{3}\Psi\times\left[(b_S + 1 - (f^C)^* - \widetilde{m}(\left(f^C\right)^*)) + (1-(f^C)^*)\left.\frac{d\widetilde{m}(f^C)}{df^C}\right|_{f^C=(f^C)^*}\right]\frac{d(f^C)^*(a^P)}{da^P}.$$

We know that $\Psi \geq 0$ and $d(f^C)^*/da^P \leq 0$, therefore, we have $(EC)'(a^P) \leq 0$ if and only if the term into brackets is positive. Replacing for $\widetilde{m}(\left(f^C\right)^* = b_S + \left(1-(f^C)^*\right)/2 - U((f^C)^*)$, we find that $(EC)'(a^P) \leq 0$ if and only if

$$(1-f^C)\left(\frac{1}{2} + \left.\frac{d\widetilde{m}(f^C)}{df^C}\right|_{f^C=(f^C)^*}\right) + U(\left(f^C\right)^*) \geq 0.$$

Since

$$\left.\frac{d\widetilde{m}(f^C)}{df^C}\right|_{f^C=(f^C)^*} = -\frac{1}{2} - \left.\frac{dU(f^C)}{df^C}\right|_{f^C=(f^C)^*},$$

we have that $(EC)'(a^P) \leq 0$ if and only if

$$(1-\left(f^C\right)^*)\left.\frac{dU(f^C)}{df^C}\right|_{f^C=(f^C)^*} - U(\left(f^C\right)^*) \leq 0.$$

We have

$$(1-f^C)\frac{dU(f^C)}{df^C} = U(f^C) + \frac{3}{2}\left(-3f^C + \frac{1}{2}D^{-1/2}\frac{dD}{df^C}\right),$$

hence,

$$(1-f^C)\frac{dU(f^C)}{df^C} - U(f^C) = \frac{-9}{2}f^C + \frac{3}{4}D^{-1/2}\frac{dD}{df^C}.$$

Finally, we have

$$\frac{dD}{df^C} = -f^C\left(12Q - 5\left(f^C\right)^2 - 6f^C + 2\right).$$

Replacing for the expression of Q, it can be show that for $t \geq \Delta q/3 + 5/4$, which is always true by assumption 1, then $12Q - 5\left(f^C\right)^2 - 6f^C + 2 \geq 0$, hence, $dD/df^C \leq 0$. It follows that

$$(1-\left(f^C\right)^*)\left.\frac{dU(f^C)}{df^C}\right|_{f^C=(f^C)^*} - U(\left(f^C\right)^*) \leq 0,$$

and that $(EC)'(a^P) \leq 0$.

7.10 Appendix J

We prove that banks' joint profits are lower if merchant 1 issues a private card than in the benchmark case. We have

$$(\Pi_I + \Pi_A)^{PC} = D_2^C((f^C)^*)((f^C)^* + m^* - c_I - c_A),$$

where D_2^C is given by C1. From Lemma 1, we know that $\widetilde{m}(f^C) \in \left[b_S + (1-f^C)/2, b_S + 3(1-f^C)/4\right]$, therefore, m^* belongs to $\left[b_S + (1-(f^C)^*)/2, b_S + 3(1-(f^C)^*)/4\right]$. Hence, we have

$$(f^C)^* + m^* - c_I - c_A \leq b_S - c_I - c_A + (3 + (f^C)^*)/4,$$

and

$$D_2^C((f^C)^*) \leq (1-(f^C)^*)\left(\frac{1}{2} + \frac{1}{6t}(-\Delta q - \frac{3}{2}\left((f^C)^*\right)^2 - \frac{1}{2} - b_S + c_M\right).$$

Since $c_M \leq c_I + c_A$, we have

$$D_2^C((f^C)^*) \leq (1-(f^C)^*)\left(\frac{1}{2} + \frac{1}{6t}(-\Delta q - \frac{3}{2}\left((f^C)^*\right)^2 - \frac{1}{2} - b_S + c_I + c_A\right).$$

Therefore, we obtain that

$$(\Pi_I + \Pi_A)^{PC} \leq h((f^C)^*),$$

where

$$\begin{aligned} h((f^C)^*) &= (1-(f^C)^*)\left(\frac{1}{2} + \frac{1}{6t}(-\Delta q - \frac{3}{2}\left((f^C)^*\right)^2 - \frac{1}{2} - b_S + c_I + c_A)\right) \\ &\quad \times \left(b_S - c_I - c_A + (3 + (f^C)^*)/4\right). \end{aligned}$$

We now show that, if $\Delta q \geq 1/2$, we have $h((f^C)^*) \leq (b_S + 1 - c_I - c_A)^2/2$ for all $f^C \in [0;1]$. Since $c_I + c_A - 1 \leq 0$, if $-\Delta q + 1/2 \leq 0$, then we have $(-\Delta q - \frac{3}{2}\left(f^C\right)^2 - \frac{1}{2} - b_S + c_I + c_A) \leq 0$. Hence,

$$\left(\frac{1}{2} + \frac{1}{6t}(-\Delta q - \frac{3}{2}\left(f^C\right)^2 - \frac{1}{2} - b_S + c_I + c_A)\right) \leq \frac{1}{2}.$$

Besides, the polynomial function $(1-f^C)\left(b_S - c_I - c_A + (3+f^C)/4\right)$ is maximal for $f^C = 2(c_I + c_A - b_S) - 1$, and its maximum is equal to $(1 + b_S - c_I - c_A)^2$. Hence,

$$h((f^C)^*) \leq \frac{(1 + b_S - c_I - c_A)^2}{2}.$$

Consequently, $(\Pi_I + \Pi_A)^{PC} \leq (\Pi_I + \Pi_A)^B$.

Cost-Based Access Pricing and Collusion with Bundling

Edmond Baranes* and Jean-Christophe Poudou†

Draft version. June 2008. Do not cite.

June 19, 2008

Abstract

This paper studies how sustainability of collusion can be influenced by both access pricing and pricing strategies of firms. We develop a model of multiproduct duopoly in vertically related industry which allows us to study how the relationship between access pricing and collusion sustainability depends on substitutability between composite goods and the direct price effect on demand.

EL Codes: L10, L50, Q40
Key words: Collusion, Access Charge, Bundling

1 Introduction

Technological convergence appears to be well underway in the telecommunications industry. Several recent studies indicate that convergence facilitates the comparison of service offerings and intensifies competition between companies. Convergence is also changing the practices adopted by firms in terms of the pricing and structure of their service offerings. To reduce the intensity of competition, firms are pursuing strategies of price discrimination between consumers. As a result, companies are multiplying their bundles or tied offers that incorporate complementary or substitutable goods. Competitive pressure and changing consumption habits are encouraging firms to market bundles of services that include telephony, internet access and television. There are several goals behind this strategy, which vary depending on the type of player offering the bundles. For instance, bundling strategies can allow entrants to win market share and incumbents to offset losses in revenues.

*Corresponding author: edmond.baranes@univ-montp1.fr. LASER, University Montpellier 1.
†LASER, University Montpellier 1.

The implications of convergence not only shape competition and pricing systems, but also lead to organizational convergence. Insofar as firms offering bundles of services do not historically come from the same markets, they do not have the same skills or core competencies, and therefore do not have access to facilities enabling them to offer these services under the same conditions. The positioning of different firms in terms of the offering of service bundles effectively depends heavily on their core competence. Strategies of extending offerings consequently do not share the same dynamic: telecom operators are looking to expand their offerings to television, whereas cable operators are adopting strategies of extending their offers to telephony and high-speed Internet access services.

Changes in the sector are raising interesting questions regarding aspects of competition. Major issues are the impacts of bundling offers both on the competitive behavior of firms and access regulation. From this point of view, the entrance of cable operators into the telecommunications markets is one of interesting example. During the last years, cable operators have upgraded their cable network infrastructure to facilitate two-way data and voice transport for cable Internet services. However, given the costs of new network deployments, cable operators could choose to extend their coverage *via* local loop unbundling rather than by building new cable. Hence, even if cable operators have a strong market power on TV market, they might buy essential facilities for broadband Internet access from telecom firms[1]. In addition, the development of Voice over Internet Protocol (VoIP) allows cable operators to enter into telephony markets and to compete hardly the incumbent who offers the telephony over PSTN. Moreover, with VoIP incumbents have to deal with competition with new upstart firms[2] offering VoIP services. Hence, anyone with a broadband connection (DSL or cable) can subscribe to a VoIP provider and make phone calls at a low rate.

This recent trend towards convergence raises interesting questions for the role of bundling on competition in telecommunications markets. For example, to what extent does competition in bundles require us to rethink the question of regulating access? Does the entrant have an incentive to use bundling to extend its market power? How does the easiness of collusion in such industries change with bundling? In this context, what is the role of access charge?

[1] For example, the UK's newly merged main cable operator ntl:Telewest has stated its intention to extend its reach via local loop unbundling of BT lines (Ofcom (2006)).

[2] Alternative operators (Yahoo!BB, Time Warner, Free, Fastweb...) or pure VoIP service providers (eBay-Skype, Google, Yahoo!...).

The recent literature on telecommunications competition and access regulation have been focuses in situations of two-way access (Armstrong (1998), Laffont, Rey and Tirole (1998a,b), Valletti and Cambini (2005)). De Bilj and Peitz (2006) build a model on that literature and analyze the emergence of VoIP networks in a PSTN environment. They focus on the effect of access regulation of PSTN networks on the adoption of VoIP. In particular, they show that higher prices for terminating access to the PSTN network make VoIP less likely to succeed.

Our paper focuses on the relationship between bundling and the feasibility of collusion when a telecom firm compete with a newcomer. The new entrant is either a firm with a full-coverage network or a provider who uses local loop unbundling to reach end-users.

During the last two decades, bundling has become an intensive research topic for Industrial Organization. Whinston (1990) clarifies the various aspects of bundling strategies and their antitrust issues. Papers initiated by Whinston (1990) have shown that the profitability of bundling results from economies of scale in the tied market. Other papers (Carbajo et al. (1990) and Seidmann (1991)) have shown that bundling may mitigate competition by inducing more differentiation. More recently, Stole (2003), Armstrong and Vickers (2007), and Thanassoulis (2007) give an interesting overview on bundling. This literature developed with legal actions against Microsoft because many economists consider that bundling has been the main driver for the development of Microsoft (Nalebuff (2004), Economides (2001)). This theoretical literature looks primarily at two cases. The first case corresponds to that of a monopolist who is threatened by an entrant and uses bundling or tying as a substitute to discrimination and to capture more consumer surplus (for instance, see Bakos and Brynjolfsson (1999)). The second case corresponds to that of an incumbent threatened by an outsider for whom bundling (or tying) is used as a means to foreclose entry Rey and Tirole (2005)). A more recent literature, in line with Matutes and Regibeau (1992), analyses competition between firms that offer bundles. In particular, Reisinger (2004) shows that the consequences of bundling are less predictable in the duopoly than in the monopoly because the traditional "sorting effect" is in balance with a "business-stealing" effect.

Although a lot of economic literature exists on bundling, this has not given rise to many papers on the relationship between bundling and collusion. Yet the existing relationship to bundling would seem to lie in the ability of firms to sustain collusion. Our framework

aims to clarify that relationship and identify the lessons to be learnt in terms of antitrust policy. It will subsequently offer relevant economic arguments regarding the justification of regulating firms' content offerings and access regulation. In an infinitely repeated game, Spector (2006) shows that the anticompetitive use of bundling is possible even in the absence of economies of scale or scope in the tied market. The mechanism from which the bundling can mitigate competition is that bundling is a tool allowing firms to shift from non-cooperation to collusion. Spector (2006) claims that if collusion is feasible in the tied market, bundling may be a profitable strategy because it may facilitate collusion.

The present paper explores how sustainability of collusion can be influenced by both access pricing and pricing strategies of firms. We develop a model of multiproduct duopoly in vertically related industry which allows us to study how the relationship between access pricing and collusion sustainability depends on substitutability between composite goods and the direct price effect on demand.

Section 2 presents the basic model where the industry is depicted and sustainability of collusion is defined. Section **??** describes the equilibrium properties for both independent and mixed-bundling pricing strategies. Section 5 explores the role of the level of access charge on collusion sustainability by exploiting relevant numerical simulations. Section **??** concludes. Proofs of Lemma and Propositions are given in an Appendix.

2 Basic model

We consider two operators indexed by i, an incumbent (operator $i = A$) and a competitor (operator $i = B$). They compete each other by offering two products X and Y. Operator A owns a complete local access network and bears a constant marginal cost normalized to 0. Since, the local loop is an essential facility, the competitor must get access from the incumbent. Let a denotes the unit access charge which firm B pays to firm A.

2.1 Consumers, demand and profits

As there are two differentiated brands of each of the two products X and Y, there are four ways to form a composite good. We denote x_i the price of product X purchased to the firm i and y_j the price of product Y purchased to the firm j and $p_{ij} = x_i + y_j$ the price of the corresponding composite good.

Consumer preferences are assumed to be given by a quadratic utility function $U(\mathbf{q})$ where $\mathbf{q} = (q_{AA}, q_{BB}, q_{AB}, q_{BA})$ represents the quantities' vector of composite goods consumed by a representative consumer[3]. More precisely

$$U(\mathbf{q}) = \sum_{h \in H} \left(\alpha_h q_h - \frac{1}{2} \beta_h q_h^2 \right) - \sum_{h,h' \in H, h \neq h'} \rho_h q_h q_{h'}$$

with $\alpha_h, \beta_h, \rho_h \geq 0$ and $H = \{AA, AB, BA, BB\}$.

For any price vector $\mathbf{p} = (p_{AA}, p_{BB}, p_{AB}, p_{BA})$,we note $D_{i,j}(\mathbf{p})$, the demand for product X purchased at the firm i and for product Y purchased at the firm j.

According to whether the composite good is sold by the same operator or not, we can consider two type of bundle: a pure bundle when both products come from the same operator and a mixed bundle when consumer mix the products.

We assume that composite goods are (imperfect) substitutes thus the demand for one good decreases with its price and increases with the other prices. This is to say that direct effects are negative, i.e. $\frac{\partial D_{ij}(\cdot)}{\partial p_{ij}} < 0, i, j = A, B$, and indirect effects are positive i.e. $\frac{\partial D_{ij}(\cdot)}{\partial p_{hk}} > 0$ where $h, k = A, B$ but $h \neq i$ and $k \neq j$. We consider that all indirect effects are of same magnitude that is $\frac{\partial D_{ij}(\cdot)}{\partial p_{hk}} = c > 0, \forall p_{hk}$. Then the parameter $c > 0$ measures substitutability between all composite goods ij. We also assume that direct effects dominate, so that demand decreases if all prices increase, $dD_{ij}(\cdot) < 0$.

In order to shed some light, we therefore restrict attention to a linear model where the demand is given by[4]:

$$D_{ij}(\mathbf{p}) = 1 - b_{ij} p_{ij} + c \sum\nolimits_{hk \neq ij} p_{hk} \tag{1}$$

The parameter $b_{ij} > 0$ measures the direct effect of price p_{ij} variations on demand D_{ij}. We further fix the following parameter configuration

$$b_{AA} = b_{AB} = b_{AB} = 1 \text{ and } b_{BB} = b$$

This assumption allows us to study an as simple as possible asymmetric structure of demands where all direct effects concerning demands addressed to firm A are normalized to 1. To ensure direct effect dominance it must true that $c < \frac{1}{3} \min\{1, b\}$.

[3]Notice that this representation assumption may be encompass by consider several consumer types with respect their degree of preference for mixing products.

[4]This implies some restrictions on parameters of the utility function $U(\mathbf{q})$.

Hence, the parameter b is not restricted to 1 in order to capture different firm specific preferences for consumers of both products. This corresponds to a situation where consumers have different preferences according to the bundle is offered by the incumbent A or the competitor B. If $b > 1$, this points out that the demand for firm B's pure bundle is more sensitive to the bundle price (i.e. p_{BB}) than other bundles. In this case (all prices being equal), consumer prefer the bundle offered by the incumbent A. A reverse interpretation applies if $b < 1$. In the sequel, we will focus our analysis in the case where demand for bundles of competitor is more sensitive to price than the incumbent's one i.e. $b > 1$.

Then we can write each operator's profit as[5]:

$$\begin{aligned}
\pi_A(\mathbf{p}, a) &= p_{AA}D_{AA}(\mathbf{p}) + x_A D_{AB}(\mathbf{p}) + y_A D_{BA}(\mathbf{p}) + \\
&\quad + a\left(D_{BB}(\mathbf{p}) + D_{AB}(\mathbf{p}) + D_{BA}(\mathbf{p})\right) \\
\pi_B(\mathbf{p}, a) &= (p_{BB} - a)\, D_{BB}(\mathbf{p}) + (x_B - a)\, D_{BA}(\mathbf{p}) + (y_B - a)\, D_{AB}(\mathbf{p})
\end{aligned}$$

At this point just notice that because of the vertical relationship between both firms, all prices remaining equals, a marginal increase of the access charge corresponds to a given marginal benefit for the incumbent A which is exactly the marginal cost incurred by the competitor B. As a result, it can be directly shown that for a given vector $\mathbf{p}$:

$$\frac{\partial \pi_A(\mathbf{p}, a)}{\partial a} = -\frac{\partial \pi_B(\mathbf{p}, a)}{\partial a} = D_{BB}(\mathbf{p}) + D_{AB}(\mathbf{p}) + D_{BA}(\mathbf{p}) > 0 \tag{2}$$

2.2 Collusion sustainability

As it is standard in the analysis of tacit collusion (Friedman, 1971), we consider an infinitely repeated Bertrand price competition game in which the punishment strategy for a given firm corresponds to trigger strategy consisting in a reversion to the a given competitive equilibrium. We denote $\pi_i^* = 0, \forall i = A, B$, the individual profit gained from a punishment strategy for firm i. We denote π_i^c individual collusion profit. The determination of π_i^c generally depends on the way the collusive agreement is reached as well as on various factors. Last, $\pi_i^{d_i}$ represents the individual profit gained from deviating from the collusive agreement.

[5] We assume that demands are not too convex in order to ensure (direct) concavity of profits that is the marginal profit of firm i is non increasing in p_{ij}.

Finally, let note δ the rate of time preference of both operators, $0 \leq \delta \leq 1$. Collusion is sustainable if, for both operators, the present discounted value of profits from being part of the cartel exceeds the present discounted value of the profits from defecting from the cartel for one stage followed by the Bertrand equilibrium profits in all subsequent stages. Thus, the incentive compatibility constraint faced by operator i gives a condition on the rate of time preference:

$$\delta \left(\pi_i^{d_i} - \pi_i^*\right) \geq \pi_i^{d_i} - \pi_i^c \tag{3}$$

Each firm is then willing to stick to the collusive price if this rate is sufficiently large. The collusive prices constitute a subgame perfect equilibrium of the infinitely repeated game if and only if $\delta \geq \max_{i=A,B} \delta_i$. As usual in the analysis of sustainability of price collusion, two effects can be distinguished : the deviation effect $\Delta_i^D = \pi_i^{d_i} - \pi_i^c$ and the punishment effect $\Delta_i^P = \pi_i^{d_i} - \pi_i^*$.

In the following, we compare the sustainability of collusion in two situations according to operators' pricing strategies. In next section, we first consider the case of Independent Pricing in which operators offer the two products X and Y separately and, second analyze the sustainability of collusion when operators choose mixed bundling strategy which consists to offer the two products separately or together with a bundle price.

As mentioned above, our interest is to show how bundling strategies affect the sustainability of collusion and whether access price makes collusion easier to sustain. In the sequel, the analysis focuses on the effect of access charge on the sustainability of collusion in a neighborhood around cost-based regulation. Hence we restrict our attention on interior equilibria which unambiguously exist[6] for access charge sufficiently close to the operating cost normalized to zero as mentioned above. In the following, we restrict our attention to two pricing strategies: Independent pricing and Mixed bundling.

3 Independent pricing

We consider the case where operators choose independent pricing. To evaluate the sustainability of collusion we have to write the incentive compatibility constraint given by (3) by calculating competition, collusion and deviation profits for each operator.

[6] This is due the linear form of the demand function we assume.

3.1 Competition, Collusion and Deviation outcomes

• **Price Competition**. First, consider the non-cooperative Nash equilibrium of the stage game, which is the usual Bertrand equilibrium in a static duopoly game. The Independent Pricing game is a collection $(x_A^*, y_A^*, x_B^*, y_B^*)$ such that:

$$\max_{(x_A, y_A)} \pi_A(x_A + y_A, x_B^* + y_B^*, x_A + y_B^*, x_B^* + y_A, a)$$
$$\max_{(x_B, y_B)} \pi_B\left(x_A^* + y_A^*, x_B + y_B, x_A^* + y_B, x_B + y_A^*, a\right)$$

The prices for composite goods are $p_{ij}^* = x_i^*(a) + y_j^*(a)$ and stand alone prices for each product are given by[7]:

$$\begin{cases} x_A^*(a) = y_A^*(a) = \dfrac{2(1 + 4b - 3c) + aA_1}{A_2} \\ x_B^*(a) = y_B^*(a) = \dfrac{10 - 6c + aA_3}{A_2} \end{cases} \tag{4}$$

We denote $D_{ij}^*(a) = D_{ij}(\mathbf{p}^*(a))$ the resulting demand for each product and $\pi_i^*(a) = \pi_i(\mathbf{p}^*(a), a)$ the competitive profit for each firm. In the neighborhood around cost-based regulation, i.e. when $a \to 0$, an interior exist for all (b, c) such that $c < \frac{1}{3}\min\{1, b\}$.[8]

• **Collusion.** When firms collude, they are assumed to behave as a cartel that maximizes total industry profits which can be written:

$$\pi_A(\mathbf{p}, a) + \pi_B(\mathbf{p}, a) = p_{AA}D_{AA}(\mathbf{p}) + p_{BB}D_{BB}(\mathbf{p}) + p_{AB}D_{AB}(\mathbf{p}) + p_{BA}D_{BA}(\mathbf{p})$$

Remark that the collusion profit is now an independent function of the access charge. Thus collusive prices will be. We have to calculate joint profit-maximizing prices (x_i^c, y_i^c) such that

$$\max_{(x_A, y_A, x_B, y_B)} \pi_A(\mathbf{p}, a) + \pi_B(\mathbf{p}, a)$$

As before, the equilibrium value of prices, demand and profits are noted p^c, D_{ij}^c, and $\pi_i^c(a) = \pi_i(p^c, a)$ the individual collusion profit.

The usual first order conditions for $\mathbf{p}(a)$ give multiple solutions. In order to focus on an unique collusive agreement, we pick a solution such that $x_A = y_A$ and $x_B = y_B$. This

[7] Where $A_1 = -(3 + 78c^2 - 27bc - 31c + 5b)$, $A_2 = 11 + 51c^2 - 40bc - 66c + 24b$, $A_3 = 75c^2 - 10bc - 46c + 7 + 6b$. One can easily check $A_2 > 0$ and $A_3 > 0$ for all admissible (b, c), while A_1 has not a constant sign.

[8] Of course if a would be significantly positive we might restrict parameters (b, c) to ensure that interior equilibria exist.

reproduces the price structure found in the competitive setting. When the cartel selects this sharing rule it annihilates competition between both mixed-bundles (AB and BA) as $p_{AB} = p_{BA}$. This implies that each pure bundle is priced twice the component price. As a result, collusion prices are[9]:

$$\begin{cases} x_A^c = y_A^c = \dfrac{b+c}{A_4} \\ x_B^c = y_B^c = \dfrac{1+c}{A_4} \end{cases} \tag{5}$$

In the neighborhood around cost-based regulation, i.e. when $a \to 0$, an interior exist for all (b, c) such that $c < \frac{1}{3}\min\{1, b\}$.[10]

Comparing (5) with (4), we can point out that even if access is priced to cost, the collusion prices may be lower than the competitive ones depending on the values of demand parameters (b and c). At first glance, it seems curious as in a more a traditional setting collusive prices are higher than competitive ones. This non classical results may come from the fact that independent pricing strategy lies composite goods' prices and hence produces strong substitutability externalities when operators choose competitively their stand alone prices. For example, when an operator decreases its price for product A to increase demand of that product, it increases in the same time demands (for mixed bundles) addressed to its rival. According to the degree of substitutability, that could create an incentive to choose high competitive prices for both operators. In this context, collusion allows to internalizes all these externalities and thus could lead to lower prices.

To ensure that collusion is feasible we must verify that each firm has an individual incentive to stick to the collusive price agreement given in (5) at least in the neighborhood around cost-based regulation. This yields us to check that, $\pi_i^c(0) - \pi_i^*(0)$ is nonnegative for all $i = A, B$. As proved in the Appendix, this is the case whenever

$$b \leq b_1(c) \tag{C.1}$$

• **Deviations.** We now turn to determine the most profitable deviation strategies for each operator. If operator i deviates from the collusive outcome in any stage of the repeated game, then it maximizes its profits in that stage, given that its competitor sets

[9] Where $A_4 = 1 - 5bc - 12c^2 + 3(b - c) > 0$ for all admissible (b, c).

[10] In collusion, one might observe that demand does not depend of the access charge. So equilibrium is interior for all values of a.

the collusive price $(x_j, y_j) = \left(x_j^c, y_j^c\right)$. Therefore, the cheating operator A or B seeks to solve:

$$\max_{(x_A, y_A)} \pi_A(x_A + y_A, x_B^c + y_B^c, x_A + y_B^c, x_B^c + y_A)$$
$$\max_{(x_B, y_B)} \pi_B\left(x_A^c + y_A^c, x_B + y_B, x_A^c + y_B, x_B + y_A^c\right)$$

where the equilibrium values are noted $p^{d_i}(a)$, $D_{ij}^{d_i}(a)$ and $\pi_i^{d_i}(a)$ when the cheating firm is i.

If operator A is the cheating firm, deviation prices are[11]:

$$x_A^{d_A}(a) = y_A^{d_A}(a) = \frac{A_5 - (1 - 5c) A_4 a}{2(3 - 5c) A_4} \tag{6}$$

Similarly if the cheating operator is B:

$$x_B^{d_B}(a) = y_B^{d_B}(a) = \frac{A_6 - a(1 + b - 4c) A_4}{2A_4(1 - 5c + 2b)} \tag{7}$$

In the neighborhood around cost-based regulation, i.e. when $a \to 0$, an interior deviation solution exist for values of (b, c) if the following condition is fulfilled

$$b \geq b_2(c) \tag{C.2}$$

The locus corresponds $b_2(c)$ to the values of parameter (b, c) such that $D_{BB}(\mathbf{p}^{d_A}) = 0$. Remark that $D_{AA}(\mathbf{p}^{d_B})$ can be zero for values of (b, c) that are excluded if (C.2) is verified.

3.2 Sustainability of Price Collusion

Substituting the relevant profits in (3) we have, for $i = A, B$

$$\delta\left(\pi_i^{d_i}(a) - \pi_i^*(a)\right) \geq \pi_i^{d_i}(a) - \pi_i^c(a) \Leftrightarrow \delta \geq \delta_i^*(a, b, c) \tag{8}$$

where $\delta_i^*(a, b, c)$ is the individual critical discount factor for firm i.

Using (8), it is possible to obtain general results about the effects of the access charge for all parameter values on the critical discount factor which is given by

$$\delta^*(a, b, c) = \max\left\{\delta_A^*(a, b, c), \delta_B^*(a, b, c)\right\}$$

If (C.1) and (C.2) are satisfied, interior solutions exist for competition, collusion and deviation outcomes and then we can state the analysis of the critical discount factor in the neighborhood around cost-based regulation in the following Proposition.

[11] Where $A_5 = 1 + 6b - 10bc - 17c^2 > 0$ and $A_6 = 2 - 7c + 5b - 17c^2 - 3bc > 0$ for all admissible (b, c).

Proposition 1 (IP) *Under conditions (C.1) and (C.2) then for cost-based access price* $(a = 0)$*,*

(i) the competitor has more incentives to deviate from the collusive agreement hence the critical discount factor is $\delta^*(0, b, c) = \delta^*_B(0, b, c)$*.*

(ii) a rise in the access charge reduces sustainability of collusion $(\frac{\partial \delta^*(0,b,c)}{\partial a} \geq 0)$ *if substitutability between composite goods is low* $(c \leq 1/7)$ *and conversely.*

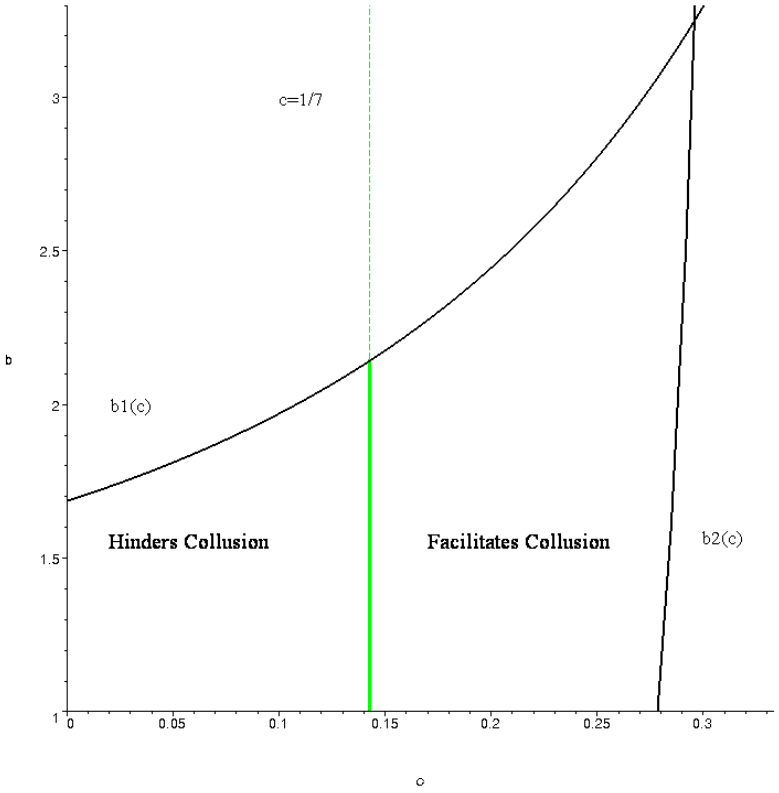
The first result *(i)* comes from the fact that gains from collusion $(\pi^c_B - \pi^*_B)$ for the newcomer (B) are lower than for the incumbent (A) since the collusive agreement internalizes demand asymmetry $(b > 1)$. Hence the critical discount factor corresponds to the one of the newcomer. The second result *(ii)* is due to the fact that both newcomer's punishment and collusion profits decreases with the access price which is a cost for this firm. However the punishment profit decreases more and more as products becomes homogeneous. Hence if substitutability between composite (c) is low, gains from collusion remains decreasing but may increase if c high enough. Notice that collusion profits (π^c_i) is less sensitive to access charge variations than punishment (competition) profits (π^*_i) since only direct effect is playing. Consequently, the variations of gains from collusion are fairly driven by those of punishment profits.

The intuition for result *(i)* is that in cost-based access regulation, the competitor does not bear any cost of access to the incumbent network as marginal cost for access is normalized to zero. This states a situation in which there is no cost asymmetry between both operators and thus describes a symmetric competition setting excepted that the demand for operator B's pure bundle is more sensitive to the bundle price than other demands $(b > 1)$. In such a case, collusive prices for the newcomer reflect more the externality created by the demand sensitivity than the incumbent collusive prices. As a result, collusion is less profitable for the newcomer.

Result *(ii)* could be explained looking at the effects of an slight increase in the access charge on the gains from collusion for the newcomer. When access charge increases, the punishment profit of the newcomer will decreases more or less depending on the degree of substitutability between composite goods. If substitutability is low, composite goods are sufficiently differentiated as competition is less intensive, hence the newcomer can report easily an increase of its access cost on its prices. When substitutability becomes higher,

composite goods becomes more homogeneous and competition more intensive. In this case, increases of the access charge create heavy burden for the newcomer as it cannot rise its prices so easily.

The following figure illustrates the results in Proposition 1, showing how, in the neighborhood around cost-based regulation, an increase of the access charge affects the sustainability of collusion.



In this figure, we see that according to the degree of substitutability between composite goods and deepness of demand asymmetry, access charge regulation (in the neighborhood of cost) may be pro-collusive or anti-collusive. In particular, supposed access charge has been set above the cost, cost-based regulation may hinder collusion when $c > \frac{1}{7}$and b low enough. This example allows us to points out a surprising results compared to the standard literature on collusion. Indeed with Proposition **??**, we obtain that more symmetry between firms might hinder collusion. This comes from the persistent effect of the demand asymmetry.

4 Mixed bundling

We consider now the case where operators choose mixed bundling. In this case, each operator i charges the price $p_{ii} = p_i$ for both products purchased as a bundle and prices x_i and y_i when consumers buy only product X or Y to operator i. Thus, the price vector writes as $\mathbf{p} = (p_A, p_B, p_{AB}, p_{BA})$. Furthermore, we assume that a given consumer who address a demand at firm i, will prefer to buy both goods as a bundle rather buy both goods as separate components if and only if

$$p_i \leq x_i + y_i \tag{C.B.}$$

We will refer to this constraint as the bundling constraint of i.

The main difference with independent pricing is that the bundle price p_i gives operator one more tool to compete each other. This allows firms to soften competition by discriminating consumers.

In the following, we give the equilibrium profits for competition, collusion and deviation.

4.1 Competition, Collusion and Deviation outcomes

• **Price Competition**. Operators choose stand alone prices for product A and B and the bundle price so as to maximize their profit. The Bertrand-Nash equilibrium of the Mixed-Bundling pricing game is then a collection $(\hat{p}_A^*, \hat{x}_A^*, \hat{y}_A^*, \hat{p}_B, \hat{x}_B, \hat{y}_B)$ such that $\forall i, j = A, B$:

$$\begin{array}{c} \max\limits_{(p_A, x_A, y_A)} \pi_A(p_A, \hat{p}_B^*, x_A + \hat{y}_B^*, \hat{x}_B^* + y_A, a) \text{ s.t. } p_A \leq x_A + y_A \\ \max\limits_{(p_B, x_B, y_B)} \pi_A(\hat{p}_A^*, p_B, \hat{x}_A^* + y_B, x_B + \hat{y}_A^*, a) \text{ s.t. } p_B \leq x_B + y_B \end{array}$$

where $\hat{\mathbf{p}}^*(a) = (\hat{p}_A^*(a), \hat{p}_A^*(a), \hat{p}_{AB}^*(a), \hat{p}_{BA}^*(a))$, with $\hat{p}_{ij}^*(a) = \hat{x}_i^*(a) + \hat{y}_j^*(a)$.

Equilibrium prices for operator A are[12]:

$$\begin{cases} \widehat{p}_a^*(a) = \dfrac{(2b - 3c^2 - 2bc + c) - 3c(2c^2 + bc - b)a}{B_1} \\ \widehat{x}_A^*(a) = \widehat{y}_A^*(a) = \dfrac{2(2b - 3c^2 - bc) - 3aB_2}{3B_1} \end{cases}$$

[12]Where $B_1 = 15c^3 - 9c^2 - 12bc + 8bc^2 + 4b$, $B_2 = 4b + 13bc^2 - 16bc - 9c^2 + 24c^3$, $B_3 = 4b + 12bc^2 - 14bc - 9c^2 + 21c^3$ and $B_4 = -6bc - 3c^2 + 2b + 4bc^2 + 9c^3$. One can check that for $i = 1, 2, 3, 4$ $B_i > 0$ for all admissible (b, c).

and for operator B:

$$\begin{cases} \widehat{p}_B^*(a) = \dfrac{(2-c-3c^2)+aB_4}{B_1} \\ \widehat{x}_B^*(a) = \widehat{y}_B^*(a) = \dfrac{2(2b-4bc+3c-3c^2)+3aB_3}{3B_1} \end{cases}$$

We denote $\hat{D}_{ij}^*(a) = D_{ij}(\hat{\mathbf{p}}^*(a))$ the resulting demand for each product and $\hat{\pi}_i^*(a) = \pi_i^*(\hat{\mathbf{p}}(a), a)$ the corresponding profit for each firm. One can easily verify that the bundling constraint (C.B.) is verified for each firm when $a = 0$. Of course it could be binding for highest values of the access charge..

• **Collusion .** When operators collude, they maximize joint-profit which writes:

$$\pi_A(\mathbf{p}, a) + \pi_B(\mathbf{p}, a) = p_A D_{AA}(\mathbf{p}) + p_B D_{BB}(\mathbf{p}) + p_{AB} D_{AB}(\mathbf{p}) + p_{BA} D_{BA}(\mathbf{p})$$

The collusion profit and collusive prices are again independent function of the access charge. The joint profit-maximizing prices (p_i^c, x_i^c, y_i^c) are such that:

$$\max_{(p_A, x_A, y_A, p_B, x_B, y_B)} \pi_A(\mathbf{p}, a) + \pi_B(\mathbf{p}, a)$$

The equilibrium value of prices, demand and profits are noted $\hat{\mathbf{p}}^c, \hat{D}_{ij}^c$, and $\hat{\pi}_i^c(a) = \pi_i(\hat{\mathbf{p}}^c, a)$ respectively. Stand alone and bundle prices are:

$$\begin{cases} \widehat{x}_i^c = \widehat{y}_i^c = \dfrac{1}{2}\widehat{p}_A^c = \dfrac{c+b}{4(b-3c^2-2bc)} \\ \widehat{p}_B^c = \dfrac{c+1}{2(b-3c^2-2bc)} \end{cases} \tag{9}$$

It should be noted that operators choose the same stand alone prices and take into account the demand direct price effect for the bundle offered by the competitor to differentiate their bundle prices. This equilibrium pricing annihilate competition between the two mixed bundles, AB and BA, by perfectly adjusting their prices, $\hat{p}_{AB}^c(a)$ and $\hat{p}_{BA}^c(a)$. The difference between bundle prices depends completely on the magnitude of parameter b with regard to 1. If the competitor's bundle demand is more sensible to its own price than the incumbent's bundle ($b > 1$), then operators charge a price $\widehat{p}_A^c$ higher than $\widehat{p}_B^c$. So, we can remark that the bundling constraint of operator A is always binding.[13] This means that when operators choose collusive prices, they don't discriminate consumers for

[13] It should be notice that when $b < 1$, the bundling constraint of operator A is always binding too.

the incumbent's products. However one can show that the bundling constraint of the competitor is always verified if $b > 1$.

Again to ensure that collusion is feasible with mixed bundling, we verify that each firm has an individual incentive to stick to the collusive price agreement given in (9) at least in the neighborhood around cost-based regulation. As proved in the Appendix, this is the case whenever

$$b \leq b_3(c) \tag{C.3}$$

• **Deviations.** To determinate deviating prices we proceed as with Independent Pricing. If the cheating operator is A its deviating prices are given by:[14]

$$\begin{aligned} \widehat{p}_A^{d_A}(a) &= \frac{(2b - 4bc - 5c^2 + c) + 2B_5 ac}{4(2c-1)B_5} \\ \widehat{x}_A^{d_A}(a) &= \widehat{y}_A^{d_A}(a) = \frac{(c + 3b - 6bc - 8c^2) + a4(3c-1)B_5}{8(2c-1)B_5} \end{aligned}$$

If the cheating firm is B, prices are[15]:

$$\widehat{p}_B^{d_A}(a) = \frac{a}{2} - \frac{(c+1)(2b - 3bc - 5c^2)}{4B_6 B_5} \text{ and } \widehat{x}_B^{d_B}(a) = \widehat{y}_B^{d_B}(a) = \frac{a}{2} - \frac{(b+c)(3b - 5bc - 8c^2)}{8B_6 B_5}$$

Again it be can verified that (C.B.) is verified for both firms. However in the neighborhood around cost-based regulation, i.e. when $a \to 0$, an interior solution exist for values of (b, c) if the following condition is fulfilled

$$b \geq b_4(c) \tag{C.4}$$

4.2 Sustainability of Price Collusion

Substituting the relevant profits in (3) we have, for $i = A, B$

$$\delta\left(\hat{\pi}_i^d(a) - \hat{\pi}_i^*(a)\right) \geq \hat{\pi}_i^d(a) - \hat{\pi}_i^c(a) \Leftrightarrow \delta \geq \hat{\delta}_i^*(a, b, c) \tag{10}$$

where $\hat{\delta}_i^*(a, b, c)$ is the individual critical discount factor for firm i with mixed bundling. Using (10), the critical discount factor which is given by $\hat{\delta}^*(a, b, c) = \max\left\{\hat{\delta}_A^*(a, b, c), \hat{\delta}_B^*(a, b, c)\right\}$.

[14] Where $B_5 = (1 - 2c)b - 3c^2 > 0$.
[15] Where $B_6 = b(1 - c) - 2c^2 > 0$.

As with independent pricing, we study how the sustainability of collusion varies with access charge. If (C.3) and (C.4) are satisfied, interior solutions exist for competition, collusion and deviation outcomes and then we can state the analysis of the critical discount factor in the neighborhood around cost-based regulation in the following Lemma.

Proposition 2 *Under conditions (C.3) and (C.4) then for cost-based access price ($a = 0$)*

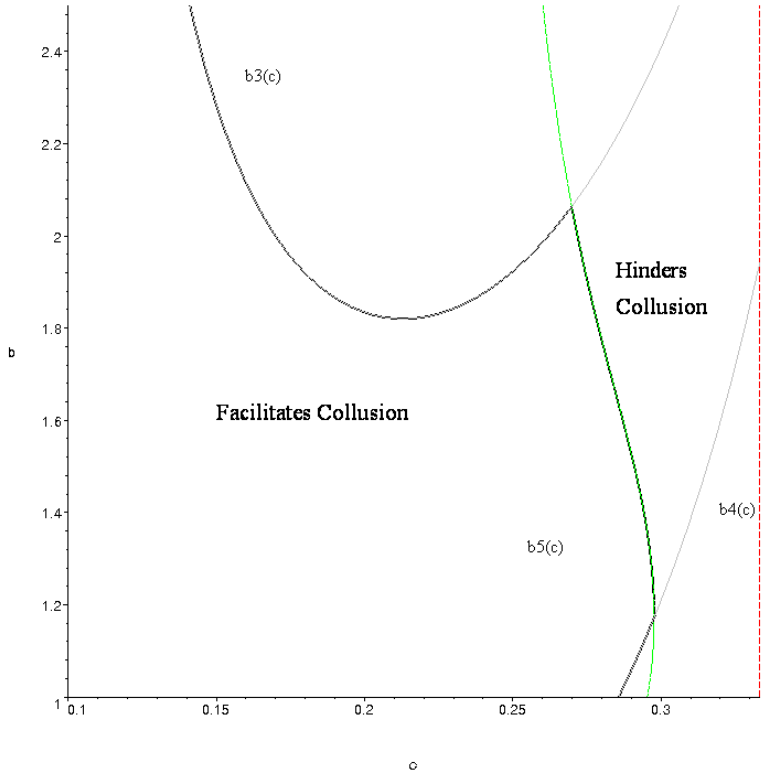
(i) ***the incumbent*** *has more incentive to deviate from the collusive agreement, hence* $\hat{\delta}^*(0,b,c) = \hat{\delta}^*_A(0,b,c)$.

(ii) it exists a locus $\tilde{c}(b)$ *such that an increase in the access charge* ***increases*** *sustainability of collusion (*$\frac{\partial \delta^*(0,b,c)}{\partial a} \leq 0$*) if* $c \leq \tilde{c}(b)$ *and conversely.*

First let us remark that with mixed bundling, direct effect of demand asymmetry is only concentrated of the bundle price of the newcomer (p_B). Consequently demands addressed to the incumbent are less concerned by the negative impact of parameter b. Then from this point of view, the internalization of demand asymmetry through the collusive agreement is less an issue for the incumbent. As a result in cost-based regulation, gains from collusion are lower for the incumbent than for the newcomer. This is the intuition for the result (i) of Proposition *2*.

Consider now an slight increase in the access charge in the neighborhood around cost-based regulation. Result *(ii)* shows that the critical factor decreases with the access charge if the degree of substitutability between composite goods is low enough relatively to demand asymmetry. This could be explained looking at the effects of gains from collusion for the incumbent. When access charge increases, the punishment profit of the incumbent will rise more or less depending on c. For the incumbent, the access charge can be viewed as a competitive advantage or at least a revenue. Hence if substitutability is relatively low, competition is less intensive and the competitive advantage coming from an increase in access charge is not so profitable for the incumbent. Then gains from collusion increase for the incumbent. When substitutability becomes higher, competition becomes more intensive. In this case, an increase in the access charge is more profitable for the incumbent in case of punishment. Then gains from collusion are now decreasing.

The following figure illustrates the results in Proposition 2, showing how, in the neighborhood around cost-based regulation, an increase of the access charge affects the sustainability of collusion.



In this figure, we see that as with Independent Pricing, access charge regulation (in the neighborhood of cost) may be pro-collusive or anti-collusive. Generally, this figure shows that, in the contrary with IP, when firm compete using Mixed Bundling strategies, access regulation may be may hinder collusion when c and b are low enough.

5 Cost-based regulation and collusion

In this section we compare, the effects of access charge regulation on collusion sustainability for both pricing regimes we have considered. From a general point of view, we see that collusion sustainability is very sensible to access charge regulation depending on pricing regimes. Access charge regulation (in the neighborhood of cost) may be pro-collusive or anti-collusive according to wether firms adopt Independent Pricing or Mixed Bundling for a given structure of demand (i.e. b and c). The following proposition presents these comparisons.

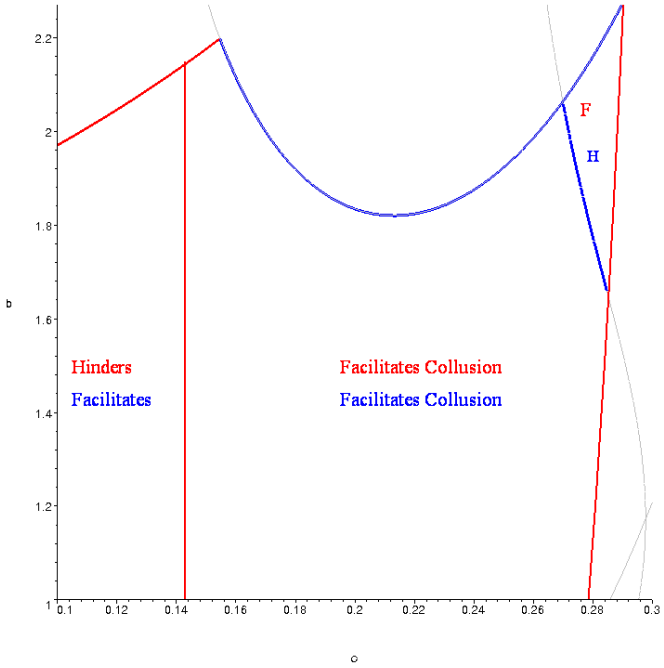
Proposition 3 *Supposed that access charge is regulated in the neighborhood around cost-based that is* $a \downarrow 0$ *then*

(i) if $1 < b \leq b_1(c)$ *and* $0 < c \leq \frac{1}{7}$*, collusion sustainability is facilitated in Mixed Bundling only*

(ii) if $b_2(c) \leq b \leq \min\{b_1(c), b_3(c)\}$ *and* $\frac{1}{7} \leq c \leq \tilde{c}$*, collusion sustainability is always*

(iii) if $b_2(c) \leq b \leq b_3(c)$ *and* $\tilde{c} \leq c < \frac{1}{3}$*, collusion sustainability is facilitated in Independent Pricing only*

For given demand asymmetry and substitutability levels, access charge regulation may facilitate or hinder collusion according to pricing regimes (Independent Pricing or Mixed Bundling). The figure illustrates Proposition 3 in case of cost-based regulation.



In particular we can see that when the degree of substitutability between composite goods and the deepness of demand asymmetry are high ($b_2(c) \leq b \leq b_3(c)$ and $\tilde{c} \leq c < \frac{1}{3}$), cost-based regulation hinders collusion sustainability with IP whereas facilitates collusion with MB. When firms coordinate on Mixed Bundling, the regulator might allow an access markup just to hinder collusion between firms.

6 Conclusion

This paper is attempt to analyze crossing effects of access charge regulation and collusive behavior when firms adopt complex pricing strategies. The strength of these effects

depends on demand substitutability and the deepness of demand asymmetry which can come from the presence of switching costs for example.

To be completed

References

[1] Armstrong, M. (1998) "Network interconnection in telecommunications", *The Economic Journal*, May, 545-564.

[2] Armstrong, M. and J. Vickers (2007) "Competitive Nonlinear Pricing and Bundling", mimeo.

[3] Bakos, Y., Brynjolfsson E.(1999) "Bundling Information Goods: Pricing, Profits, and Efficiency", *Management Science*, vol. 45, N°12, December, pp. 1613-1630.

[4] Carbajo J., D. deMeza, and D.J. Seidmann (1990) "A Strategic Motivation for Commodity Bundling" *Journal of Industrial Economics*, 38, 283-298.

[5] Chang, M.-H. (1991) "The effects of product differentiation on collusive pricing", *International Journal of Industrial Organization*, 3, pp. 453-470.

[6] De Bijl, P.W.J. and M. Peitz (2006) "Access Regulation and the Adoption of VoIP", *mimeo,* may.

[7] Dessein, W. (2004) "Network competition with heterogeneous customers and calling patterns", *Information Economics and Policy*, Vol. 16, 3, 323-345.

[8] Economides N. (2001) "The Microsoft antitrust case", *Journal of Industry, Competition and Trade: From Theory to Policy*, vol. 1, no. 1, pp. 7-39.

[9] Friedman, J.W.(1971) "A non-cooperative equilibrium for supergames", *Review of Economic Studies,* 38(1), 1-12.

[10] Laffont, J.J., P. Rey, J. Tirole (1998a) "Network competition: I. Overview and nondiscriminatory pricing", *Rand Journal of Economics*, vol. 29, n°1, spring 1998, 1-37.

[11] Laffont, J.J., P. Rey, J. Tirole (1998b) "Network competition: II. Price discrimination", *Rand Journal of Economics*, vol. 29, n°1, spring 1998, 38-56.

[12] Matutes C. and P. Regibeau (1992) "Compatibility and Bundling of Complementary Goods in Duopoly", J*ournal of Industrial Economics*, 40, pp. 37-54.

[13] Nalebuff B. (2000) "Competing against bundles", WP7, Yale School of Mangement.

[14] Ofcom (2006) The Communications Market 2006, August.

[15] Reisinger M. (2004) "The Effects of Product Bundling in Duopoly", *Discussion Paper* 2004-26, University of Munich.

[16] Rey P. and J. Tirole (2005) "A Primer on Foreclosure", forthcoming, *Handbook of Industrial Organization*, vol. 3, ed. By M. Armstrong and R.H. Porter, Noth Holland.

[17] Seidmann D.J. (1991) "Bundling as a Facilitating Device: A Reinterpretation of Leverage Theory" *Economica*, 58, 491-499.

[18] Spector D. (2007) "Bundling, tying, and collusion", *International Journal of Industrial Organization*, (25)3, 575-581.

[19] Stole L.A. (2003) "Price Discrimination and Imperfect Competition", Working Paper, University of Chicago, forthcoming in: *The Handbook of Industrial Organization*, vol. 3, ed. By M. Armstrong and R.H. Porter, Noth Holland.

[20] Thanassoulis J. (2007), "Competitive Mixed Bundling and Consumer Surplus", *Journal of Economics & Management Strategy*, 16(2), 437–467.

[21] Valletti, T.M. and C. Cambini, (2005) "Investment and network competition", *Rand Journal of Economics*, vol 36, n°2, 446-467.

[22] Whinston M.D. (1990) "Tying, Foreclosure and Exclusion", *American Economic Review*, 80: pp. 837-859.

Appendices

• **Derivation of Condition (C.1)**: This restriction comes from the inequality $\pi_B^c(0) - \pi_B^*(0) \geq 0$ with

$$\pi_B^c(0) - \pi_B^*(0) = 2\,\frac{(7\,c-1)^2\left[(40\,c-24)\,b^2+(72\,c-84\,c^2+28)\,b+9\,c^3+63\,c^2+21-117\,c\right]}{A_4\;A_2^2}$$

Since $A_4 > 0$ the sign of $\pi_B^c(0) - \pi_B^*(0)$ is such that

$$\text{sign}(\pi_B^c(0) - \pi_B^*(0)) = \text{sign}\left[(40\,c - 24)\,b^2 + \left(72\,c - 84\,c^2 + 28\right)b + 9\,c^3 + 63\,c^2 + 21 - 117\,c\right]$$

then $\pi_B^c(0) - \pi_B^*(0) \geq 0$ if $b \leq b_1(c) = \frac{-18\,c+21\,c^2-7+(5-3c)\sqrt{39\,c^2-18\,c+7}}{4(5\,c-3)}$. Studying $b_1(c)$ for $c \in [0, \frac{1}{3}]$, shows it is an strictly increasing function from $\frac{1}{12}\left(7+\sqrt{175}\right) \simeq 1.686$ to $2 + \frac{768}{16} \simeq 3.732$. Notice that for $b > b_1(c)$ then $\delta^B(0,b,c) = 1$ and no collusion is sustainable in the industry. ■

• **Derivation of Condition (C.2)**. This restriction comes from the fact that $D_{BB}^{d_A}(a)$ can be zero for some values of (b,c) even if a is around 0. Indeed this equilibrium demand writes

$$D_{BB}^{d_A}(a) = \frac{(1-3\,c)(3-5\,c)\,b + 16\,c^3 - 25\,c^2 - 6\,c + 3}{(3-5c)\,A_4} + \frac{2(5c-1)c}{(3-5c)}a$$

Hence $D_{BB}^{d_A}(0) \geq 0$ if

$$b \geq b_2(c) = \max\{1, \frac{25\,c^2 - 16\,c^3 + 6\,c - 3}{(3-5\,c)(1-3c)}\}$$

where $b_2(c) = 1$ for $c \in [0, \frac{1}{16}\left(13-\sqrt{73}\right)]$ and $b_2(c)$ is strictly increasing towards ∞ for $c \in [\frac{1}{16}\left(13-\sqrt{73}\right), \frac{1}{3}[$. One could notice that $D_{AA}^{d_B}(0)$ can also be zero but for lower values of b since $D_{AA}^{d_B}(0) \geq 0$ if $b \geq \beta_2(c) = \max\{1, \frac{(1+c)(11c-3)+\sqrt{505c^4-154c^2+72\,c^3+24\,c+1}}{4(1-3c)}\}$ with $\beta_2\left(\frac{13}{16} - \frac{\sqrt{73}}{16}\right) = 1$, hence $\beta_2(c) < b_2(c)$ for $c > \frac{1}{16}\left(13-\sqrt{73}\right)$ ■

• **Proof of Lemma 1**: To prove point *(i)*, one can see that $5\,c - 2\,b - 1 < 0$ and $5\,bc + 12\,c^2 + 3\,c - 3\,b - 1 < 0$ for $c \in [0, \frac{1}{3}[$ and $b > 1$, then

$$\Delta_B^{D*} - \Delta_A^{D*} = \frac{(7\,c-1)^2(b-1)}{2(3-5\,c)(5\,c-2\,b-1)(5\,bc+12\,c^2+3\,c-3\,b-1)} \geq 0$$

Moreover $51\,c^2 - 66\,c - 40\,bc + 11 + 24\,b > 0$ and $249\,c^2 - 80\,bc - 174\,c + 48\,b + 2 > 0$ for $c \in [0, \frac{1}{3}[$ and $b > 1$, then

$$\Delta_B^{P*} - \Delta_A^{P*} = -\frac{(b-1)(7\,c-1)^2(249\,c^2 - 80\,bc - 174\,c + 48\,b + 25)}{2(3-5\,c)(51\,c^2 - 66\,c - 40\,bc + 11 + 24\,b)(5\,c-2\,b-1)(5\,bc+12\,c^2+3\,c-3\,b-1)} \leq$$

As a result $\delta_B(0,b,c) \geq \delta_A(a,b,c)$ then $\delta^*(0,b,c) = \delta_B(0,b,c)$. This complete the proof of point (i) of the Lemma.

To prove point *(ii)*, we calculate the derivative $\delta^*(a,b,c)$ with respect to a around $a = 0$, that is

$$\frac{\partial \delta^*(0,b,c)}{\partial a} = \frac{(c+b)(51\,c^2 - 66\,c - 40\,bc + 11 + 24\,b)^2(4\,bc - 2\,b - 1 + 2\,c + 9\,c^2)(5\,c-2\,b-1)(5\,bc+}{4(1587\,c^4 - 1929\,c^3 - 515\,bc^3 - 657\,bc^2 - 533\,c^2 - 240\,b^2c^2}$$

One can see that for $c \in [0, \frac{1}{3}[$ and $b > 1$, $4\,bc - 2\,b - 1 + 2\,c + 9\,c^2 < 0$ and $-64(3 - 5c)(1 + c)b^2 + ... < 0$ hence $\text{sign}(\frac{\partial \delta^*(0,b,c)}{\partial a}) = \text{sign}(1 - 7\,c) \lesseqgtr 0$ iff $c \gtreqless \frac{1}{7}$. ■

• **Derivation of Condition (C.3)**: This restriction comes from the inequality $\hat{\pi}_A^c(0) - \hat{\pi}_A^*(0) \geq 0$. Since $3c^2 + (2c - 1)\,b < 0$ then

$$\begin{aligned} \hat{\pi}_A^c(0) - \hat{\pi}_A^*(0) &= -\frac{C(b,c)}{18\,(3c^2 + (2c-1)\,b)\,B_1^2} \\ \text{sign}(\hat{\pi}_A^c(0) - \hat{\pi}_A^*(0)) &= \text{sign}C(b,c) \Leftrightarrow \text{sign}(\pi_B^c(0) - \pi_B^*(0)) \geq 0 \text{ if } b \leq b_3(c) \end{aligned}$$

where

$$\begin{aligned} C(b,c) &= 8\,(11\,c-1)\,(c-1)\,(2\,c-1)^2\,b^3 + 24\,c\,(2\,c-1)\left(28\,c^3 - 31\,c^2 + 10\,c - 3\right)b^2 + \\ &\quad +9\,c^2\left(-2 - 32\,c - 234\,c^3 + 161\,c^4 + 151\,c^2\right)b + 27\,c^4\left(-40\,c^2 + 11\,c^3 + 2 + 15\,c\right) \end{aligned}$$

Using analytical Cardan method one can see that the discriminant Δ of the canonical form of the equation $C(b,c) = 0$ is positive for all $c \in [c_0, \frac{1}{3}]$, and negative on $[0, c_0[$, where $c_0 \simeq 0.0589$. The discriminant Δ is

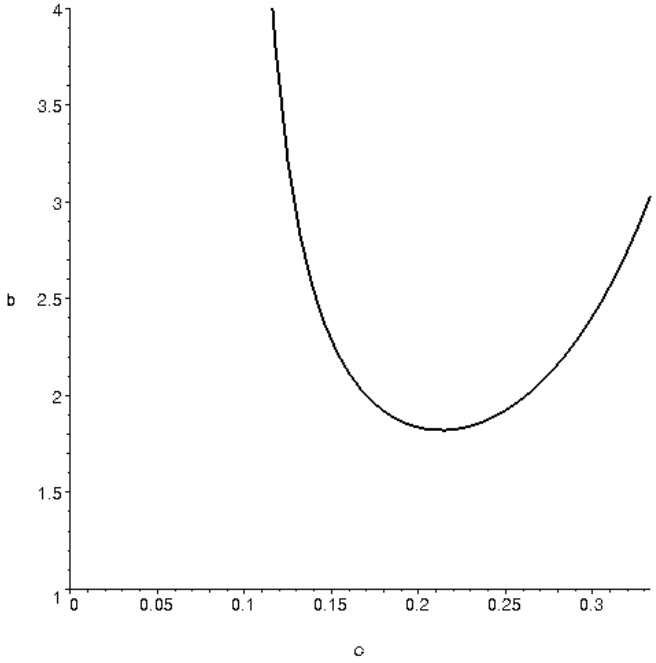
$$\Delta = \frac{27}{128}\frac{(6571\,c^5 - 16303\,c^4 + 12154\,c^3 - 3794\,c^2 + 524\,c - 20)\,(3\,c-2)^2\,(c+1)^7\,c^6}{(1-2\,c)^6\,(1-c)^4\,(11\,c-1)^4}$$

Hence this cubic equation has only one real root greater than 1 if $c \geq c_0$ (see figure below) we denote $b_3(c)$ and which is given by

$$b_3(c) = \sqrt[3]{\frac{-q + \sqrt{\Delta}}{2}} + \sqrt[3]{\frac{-q - \sqrt{\Delta}}{2}} - \frac{c\,(28\,c^3 - 31\,c^2 + 10\,c - 3)}{(11\,c-1)\,(1-c)\,(1-2\,c)}$$

where

$$q = -\frac{(11593\,c^6 - 39237\,c^5 + 52671\,c^4 - 36683\,c^3 + 14193\,c^2 - 2844\,c + 243)\,(c+1)^3\,c^3}{4\,(1-2\,c)^3\,(1-c)^3\,(11\,c-1)^3}$$



• **Derivation of Condition (C.4)**. This restriction comes from the fact that $D_{AA}^{d_B}(0)$ can be zero for some values of (b,c), this equilibrium demand writes

$$D_{AA}^{d_B}(0) = \frac{(c^2 - 5\,c + 2)\,b^2 + 2\,c^2\,(3\,c - 5)\,b + c^3\,(-1 + 7\,c)}{4\,(2\,c^2 + bc - b)\,(3\,c^2 + 2\,bc - b)}$$

Since $3c^2 + 2bc - b < 0$ and $2\,c^2 + bc - b < 0$ then $D_{BB}^{d_A}(0) \geq 0$ if the following quadratic inequation of b is solved: $(c^2 - 5\,c + 2)\,b^2 + 2\,c^2\,(3\,c - 5)\,b + c^3\,(-1 + 7\,c) \geq 0$, that is if

$$b \geq b_4(c) = \max\{1, \frac{c\,(3c - 5) - (1 + c)\,\sqrt{2(1 + c)c}}{2 - 5\,c + c^2}\}$$

where $b_4(c) > 1$ for $c \in]\frac{2}{7}, \frac{1}{3}]$ and a nondecreasing function of c. ■

• **Proof of Lemma 2**: To prove point *(i)*, one can see that $2\,c^2 + bc - b < 0$ and $3\,c^2 + 2\,bc - b < 0$ for $c \in [0, \frac{1}{3}[$ and $b > 1$, then

$$\hat{\Delta}_B^{D*} - \hat{\Delta}_A^{D*} = -\frac{(1 - b)\,(4bc - 2\,b + 7\,c^2 + c)\,c}{16\,(1 - 2\,c)\,(2c^2 + bc - b)\,(3c^2 + 2\,bc - b)} \leq 0$$

Moreover $4bc - 2\,b + 7\,c^2 + c < 0$ if condition (C.3) if verified that if $b \geq b_3(c)$.

One can see $15\,c^3 - 9\,c^2 + 8\,bc^2 - 12\,bc + 4\,b > 0$ and $24b^2\,(c - 1)\,(2c - 1)^2 + 2bc\,(2c - 1)\,(-59c + 89c^2 - 4$ $3c^3\,(-40c + 109c^2 - 5) < 0$ for $b \geq b_3(c)$ then

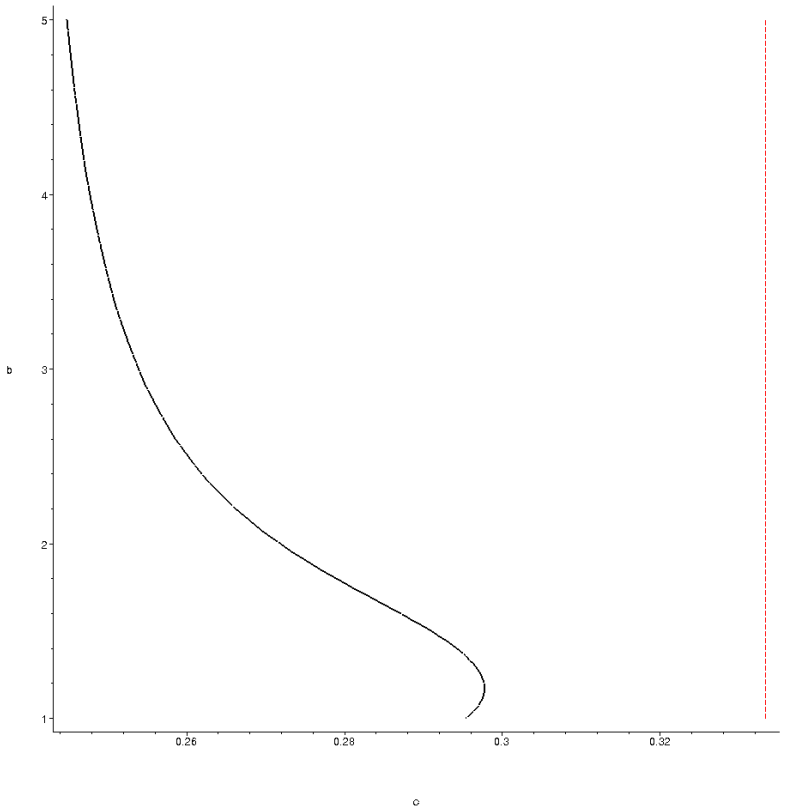
$$\hat{\Delta}_B^{P*} - \hat{\Delta}_A^{P*} = \frac{c\,(b - 1)\,(-192\,b^2c^2 + 120\,b^2c + 96\,b^2c^3 - 24\,b^2 + 356\,bc^4 - 414\,bc^3 + 102\,bc^2 + 8\,bc + 327\,c^5}{16\,(2\,c - 1)\,(2\,c^2 + bc - b)\,(3\,c^2 + 2\,bc - b)\,(15\,c^3 - 9\,c^2 + 8\,bc^2 - 12\,bc + 4\,b)}$$

As a result $\hat{\delta}_A(0,b,c) \geq \hat{\delta}_B(0,b,c)$ then $\hat{\delta}^*(0,b,c) = \hat{\delta}_A(0,b,c)$. This complete the proof of point (i) of the Lemma.

To prove point *(ii)*, we calculate the derivative $\hat{\delta}^*(a,b,c)$ with respect to a around $a=0$, i.e.

$$\frac{\partial \hat{\delta}^*(0,b,c)}{\partial a} = -\frac{144\left(3\,c^2+2\,bc-b\right)^2\left(15\,c^3-9\,c^2+8\,bc^2-12\,bc+4\,b\right)^2\left(1-2\,c\right)}{Y^2} D(b,c)$$

Then it turns that the sign of $\frac{\partial \hat{\delta}^*(0,b,c)}{\partial a}$ is the one $-D(b,c)$. The expression $D(b,c)$ is a quartic parametric function of b for we represent the admissible root (in the sense $b \geq 1$ and real) using a Maple software:



Denoting this locus $b_5(c)$, we see that it splits the admissible space of parameter's values $\left(\left[0,\frac{1}{3}\right]\times[1,\infty)\right)$ in two parts. Testing for an admissible couple $(b,c)=(1,0)$ for which $b=1<b_5(0)\to\infty$ leads to $D(1,0)=352$ so $\frac{\partial \hat{\delta}^*(0,b,c)}{\partial a}<0$ and picking another arbitrary admissible couple $(b,c)=\left(3,\frac{1}{3}\right)$ for which $b=3>b_5\left(\frac{1}{3}\right)\simeq 0.254$ implies $D\left(3,\frac{1}{3}\right)=-125.18$ so $\frac{\partial \hat{\delta}^*(0,b,c)}{\partial a}>0$. Then if $b\leq b_5(c)$ then $D(b,c)\geq 0$ and if $b>b_5(c)$ then $D(b,c)>0$. Moreover since $b_5(c)$ is strictly decreasing (for admissible values of b and c) we denote $\tilde{c}(b)=b_5^{-1}(b)$. Then condition $b\leq b_5(c)$ is equivalent to $c\leq\tilde{c}(b)$. ■

Triple play vs Software voice in France

Cecere Grazia
Université Paris Sud 11, Laboratoire ADIS, 54, Bl. Desgranges, 92331 Sceaux cedex, France
University of Torino, Dipartimento "S. Cognetti De Martiis", Via Po, 53. I, 10124 Torino, Italy

Preliminary version September 2008

Abstract

On the demand side, great attention has been given to the diffusion of ICT services and devices among consumers and households. But what determines consumers' choices when competing services appear on the market? Is there an effect of complementarity and substitution? This study applies the economics of innovation literature to an analysis of the determinants of the adoption of two services that enable unlimited voice communications (offered by VoIP providers), i.e. software voice and IP network voice services. We employ a bivariate probit model, which allows us to take account of the possible decision to adopt both services. The empirical investigation is based on French micro level data collected in 2005 for monitoring the ICT usage by households and individuals. The data were collected by INSEE (French Statistic Office). Results entail some major policy and regulatory implications.

Key words: innovation diffusion among consumers, bivariate probit, complementarity vs. substitution effect

Introduction

The innovation diffusion process is the sum of complex interactions among firms, consumers, institutions and industries (Mowery, Nelson, 1999 ; Van den Ende, Dolfsma, 2005). The speed of these processes is determined by "the characteristics of the product's technology" and the consumer preferences (Klepper, Graddy, 1990: 35). In network industries, the choice of consumers' adoption decision can influence the rate and direction of diffusion process. We aim to investigate on the demand side decision to adopt unlimited voice services on a sample of French individuals. On the supply side, at the time of the survey in 2005, there are mainly two kinds of services offering reduced or unlimited voice communication. This raises economic questions hinging on the effect of complementarities and substitution among these two services which have policy and regulatory implications.

The telecommunication sector is characterized by the rapid pace of technological changes. The introduction of VoIP (Voice over Internet Protocol) technology has had a direct impact on consumers in providing voice communications that are either completely free or at very reduced prices. These voice services are of two types: software voice (SOFTWAREVOICE)[1] and IP network services (TRIPLAY)[2]. What are factors determining the adoption of these two different services? On the supply side, could these services replace the bill-and-keep tariff system of pricing?

[1] In the article both software voice and SOFTWAREVOICE expression would be used as well as IP network voice services and TRIPLAY.

[2] At the time of the survey, France Telecom, the incumbent, had started to experiment the offer of unlimited voice communication bundled with internet and video to its customers

We drawn on the literature on innovation diffusion coupled with the literature on telecommunication industry sector specific features to identify the determinants of consumers' patterns of adoption of the voice service offered by VoIP providers. We use French cross-section micro level data based on the survey built in 2005 by the French statistics office - INSEE. As concerned the econometric model, we deal with discrete variable setting, the probabilities of adopting Software voice and of subscribing to IP network voice are simultaneously determined by a bivariate probit model, which allows for correlation of the unobserved effects and errors. For deeply testing the adoption patterns of Software voice, we estimate univarite probit explaining the adoption of these services taking in consideration internet related variables.

Our empirical investigation tests three hypotheses. (1) The more technophile individuals have more probabilities to adopt software voice services in respect to IP network voice subscribers. As, software voice application can be included among the internet related activities. (2) We identify the variables influencing the effect of complementarity or/and substitution adoption of these two services - software voice and IP network voice. Indeed, software voice requires some internet related competences, whereas subscribing to an IP network service requires no specific capabilities. (3) The influence of geographical place of residence has been tested as it captures both the exchange and density of local information and the impact of IP network service which is endogenously determined. At the time of the survey these services were accessible only in the most populated areas, which have implications at policy perspective e.g. the diffusion of the broadband. Also, among European countries, France is one of the most advanced in terms of penetration of broadband and related services.

Section 1 presents the empirical questions and provides some data showing the situation of French telecommunication sector in respect to VoIP diffusion and to the broadband penetration. Section 2 reviews the literature on economics of innovation for analysing the patterns of innovation adoption among consumers. Section 3 describes the data and Section 4 presents the econometric model and the results. Section 5 shows the results of the estimation of the Software voice adoption choice. Section 7 the underlines the main limits. Conclusion follows.

1. Telecommunication sector and effect of substitution and complementarities

Our study combines the literature on innovation diffusion with the analysis of telecommunication industry sector-specify features. Both software voice and IP network service providers are part of the info-communication sector, which is an internet-based industry (Fransman, 2003, Krafft, 2003, 2004) exploiting ICT as a GPT; it qualifies as a « demand driven » (Krafft, 2004) industry. Consumer choices can determine the direction of self reinforcing process and the dominant position in the market.

The introduction of VoIP has enabled to reduce cost or offer unlimited free voice traffic and value added services. Two main models of VoIP applications are identified: SOFTWAREVOICE (e.g. Skype, JaJah) and TRIPLAY. In France, at the time of the survey IP network voice providers (new entrants to the telecommunication sector), such as Iliad/Free, Neuf Cegetel (Internet Service Providers-ISP) were offering unlimited national and international calls to consumers within subscription fees bundled in the form of triple (data, voice and video) and double play (data and voice) with other value-added services such as email. The price of these services range among the 30 euros per month. These services require

a broadband network. As at the time of the survey, the incumbent, France Telecom, has started to offer comparable set of services as an experimental offering. It is possible to assume that individual subscribing to IP network service choose new entrant offer. In other words, consumers analyzed could be defined innovators or early adopters.

Software voice providers allow consumers to make calls using the internet with unlimited free calls from PC to PC, the necessary condition of using this service is to have a PC connected to internet. Consumers download free software, which enables them to do instant messaging, send short messages and send files over the internet.

Table 1 presents the statistics showing the evolution of households' subscriptions of different telecommunication services in France. The last row in the table indicates the number (in millions) of VoIP subscriptions related to IP network voice services. The data do not include communication through software voice services. It emerges that subscription to VoIP are increasing thus, the new operators might challenge the incumbents as they propose more advantageous offer to consumers. As already mentioned, the diffusion of the broadband is a necessary condition for access to unlimited calls offered by ISPs. Table 2 presents broadband diffusion for some European countries.

Table 1: French subscriptions to telecommunication services

Millions line	2002	2003	2004	2005	2006	Evol.
Nb of subscription in the last period	34,124	33,913	34,541	36,498	38,168	4,6%
Subscription to analogical network	28,980	28,673	28,502	27,969	26,477	-5,3%
Subscription to numeric network	5,084	5,176	5,038	5,002	4,872	2,6%
Subscription to cable	0,058	0,060	0,069	0,135	0,211	55,7%
Subscription to VoIP	-	-	0,931	3,392	6,608	94,8%

Source : ARCEP (Survey from 1998 to 2005): p. 177

Table 2: Diffusion of broadband in Europe, December 2006

	Broadband access	Penetration rate	Cable	Broadband offered by incumbents	Broadband offered on the whole market	Unbundling lines	Bitstream lines
France	10 819 301	18%	600 000	4 873 263	5 514 106	3 513 133	2 00 973
Germany	11 666 2002	14%	284 250	6 500 00	6 444 300	3 543 000	2 901 300
Italy	7 381 612	13%	0	4 928 000	2 507 122	1 432 122	1 075 000
Spain	5 362 119	13%	1 169 666	3 084 555	1 184 802	559 563	625 239
United Kingdom	11 051 967	19%	2 870 354	2 584 000	6 362 802	838 379	5 459 000
Holland	4 360 121	27%	1 550 00	1 970 690	796 560	796 560	0
Total/ Mean of the 25 European countries	66 548 642	15%	10 037 901	32 993 729	25 334 849	12 011 886	13 322 963

Source: Arcep website ECTA (European Competitive Telecommunication Association) data, December 2006 [3]

1.2 Substitution vs. complementarity

When, there are two competing patterns of innovation, the substitution vs. complementary effect on the demand side is crucial for investigating the outcome of competition. In microeconomics, the substitution effect is computed by calculating the cross elasticity of the prices of two goods. Increasing the consumption of a good A decreases the consumption of the substitute good B. In knowledge based industries, firms operate within more complex systems, and the competition dynamics are different from traditional industry. Here, competition is based on performance (Pleatsikas, Teece, 2001) and creativity, rather then price reduction, and even more so when the services offered are free. Thus, when prices cannot be computed or services are free, the effect of complemetarity/substitution has to be calculated using quantities rather than cross price elasticity using a discrete choice setting (Gentzkow, 2007).

[3] http://www.arcep.fr/index.php?id=9184#

The concepts of substitutability and complementarities in the telecommunication sector were tackled essentially to study the substitution between fixed and mobile telephone access in developed (Rodini et al., 2003) and developing countries (Hamilton, 2003; Garbacz, Thompson, 2007), while Duffy-Deno (2001) looked at the complementarity effect of the second telephone line diffusion in the US. The notion of substitution and complementarity should be considered in broad terms; it overlaps with economic and technical concepts - in other words, (i) the functionalities of goods, (ii) consumers' perceptions and (iii) barriers to substitution.

(i) The substitution functionality includes quality of service and technical features. Technically, both types of voice services offer the possibility to receive and to make calls (Corrocher, 2003). The quality of the communication with software voice transmission initially was poor, but this is no longer the case. The quality of TRIPLAY transmissions was always better and is now almost problems free.

(ii) The consumers' perceptions of the technology (Mindel, Sicker, 2006) identify their thinking about and perception of reliability of the services. This concept is used into competition framework for determining the relevant market as it can determine the conditions for product differentiation (Andreosso, Jacobson, 2005). Consumers have different perceptions of the services provided as their functionality substitution is not perfect. On the one hand, TRIPLAY voice communication has characteristics similar to the classic telephone as individuals have to use a classical handset for making calls[4]. On the other hand, consumers using SOFTWAREVOICE need some ICT competences, i.e. the know how to use the service

[4] The individuals have to install a home gateway for the connection but these companies give engineering support.

reduces the “cognitive dissonance” (Klemperer, 1995) of individuals toward the technology. This leads an obstacle of the adoption and on the Klemperer’s assumption, it increases the switching cost of adoption.

(iii) The barriers to substitution lead to geographical coverage and competences related. As concerned the geographical coverage, users would have access to software voice providers through any internet connection (in the home or in a public place, although some administrations have tried to ban use of this software application). Triple play access is conditional on the geographical penetration of the broadband. The regulatory agencies do not include software service providers among the telecommunication providers. As concerned the competences related barriers, the SOFTWAREVOICE service is internet competence demanding while the TRIPLAY service does not require nay particular competences.

The econometric approach informs about the sustainability of these two innovative services. All characteristics hold, two services could be also defined as imperfect substitute (Greenstein, Mazzeo, 2006): they can satisfy the same needs, but the conditions of adoption are different. An alternative view is that these two voice communication services are complements. Both services give the possibility to make and receive calls and they also satisfy other diverse needs. For IP network service adopters, the software voice applications could be considered as a complements as it enables instant messaging and transferring of files and video conferencing. While the Software voice adopters might consider IP network voice as substitute as it could enable voice communication.

2. The innovation diffusion and adoption.

This study draws on the literature economics of innovation coupled with that on network sector-specific industries, an approach adopted by Majumdar, Venkataraman (1998), Constantiou et al. (2008) and Michalakelis et al. (2008). The empirical literature of economics of innovation leading to consumers analysis has concentrated on the patterns of ICT adoption and usages. The main topics emerging are: (1) the diffusion of ICT devices such as the PC (Demoussis, Giannakopoulos, 2006), the choice of mobile handset mobile phone; (2) the influence of Internet applications in changing consumers' behaviour (Hong, 2007), (3) the patterns of adoption vs. usage of internet (Goldfard, Prince, 2008); (4) the effect of substitution and complementarities between mobile and fixed telephone (see section 1.2); (5) in terms of the telecommunication service *stricto sensu,* the literature investigates the rate of penetration of fixed lines, mobile phones and broadband connections and providing forecasting analysis (Garbacz, Thompson, 2007).

Adoption takes place when consumers or firms choose/purchase products, services and new organizational structures that are shaped or supplied by other firms, consumers or organizations (Antonelli, 2006). We consider the adoption decision as part of the diffusion process; aggregate adoption enables the diffusion process[5]. The new services or goods can meet the needs of users or they can create new needs. In our empirical study, we have information on the innovation adoption, thus we try to identify the condition enabling diffusion.

[5] As we underline in the next sections, we dispose of cross section and we do not have information about the dynamic of adoption, thus we can only a static model.

In the literature on innovation diffusion consumers are categorized as: innovators, early adopters, late majority, early majority or laggards (Rogers, 1983: 246, 247). What mainly distinguishes these categories is the time lag in adoption, the propensity of consumers to adopt the technology in term of the capabilities required and thus the different degrees of marginal utility they gain (Swann, 2002). This literature has developed a framework that can be used for the purposes of forecasting (Bass, 1969, 1980). Other methods of forecasting, such as Gompertz's curve, exploit demand elasticity due to price variations (Fildes, Kumar, 2002; Robertson et al., 2007). The decision to adopt ICT could also be driven by non-capital costs of adoption (Astebro, 2002:673) such as ability and confidence in the technology.

The literature on innovation economics combines different "theoretical approaches into a single framework" (Karshenas, Stoneman, 1993; Faria et al., 2002: 570), namely the epidemic model, the rank effects model (Battisti, Stoneman, 2005), the model focusing on network effects as the driving force of innovation diffusion and the order effect model. We refer to the epidemic and network effects models.

Epidemic model

The epidemic model examines the speed and spread of information from users to non-users (Mansfield, 1961; Geroski, 2000; Bochet, Brossard, 2007) assuming that individuals are homogenous. The epidemic model is based on the mathematical approach of the contagion model. The adoption decision is influenced by social constraints and the satisfaction of new needs. The information exchange is based on word of mouth exchanges (local information), which is also described as local spill-over (Le Guel et al., 2005; Roux, Galliano, 2007), and on common information sources (global information such as broadcast). The spread of

information on the technology drives innovation diffusion, hence more information about the innovation reduces the degree of uncertainty. A large base of early adopters increases the probability that adopters will contact non-adopters. These effects can be captured by region/area inhabitants' density and by data on information exchange within the social network of consumers (Steyer, Zimmerman, 2004), in other word it aims to capture the proximity among individuals (Torre, Rallet, 2005).

Network effects

The network effects imply that the importance and the value of a network is determined by the number of participants connected (Economides 1995, Shapiro and Varian, 1998); hence, the value increases with the number of participants. Individuals adopt/buy the service because they want to joint the network. Individuals are supposed to be heterogeneous. The network effect consists of the positive connection between "adopters" and the dimension of the network (Church, Ware, 1998)[6]. This creates the conditions for a self reinforcing process (Katz, Shapiro, 1986) and increasing returns to adoption (Arthur, 1989). Consumers might be lock-in by services or goods. (Geroski, 2000: 619). If consumers are locked into a technology, changes might incur switching costs (David, 1985).

We need to distinguish between the general definitions of network effects where it does not matter who are the members of the network (Majumdar, Venkataraman 1998), and the social network which represents the group of peers that belong to the network (Rolfhs, 1974; Birke, Swann, 2006). Goolsbee and Klenow (2002) analyse also the concept of local network effect,

[6] Ibid. p. 228

where geographical proximity influence the adoption of network services and goods. In the case of software voice service consumers are encouraged to joint network when relatives or interest groups are members of that network.

Effect of first order innovation adoption

GPT (General Purpose Technology) innovations enable various applications, in diverse sectors, and create new applications opportunities (Bresnahan, Trajtenberg, 1995). GPT consumers are customized to the application of GPT (Steinmueller, 2006); these adopters of first generation innovations have the capabilities to adopt second generation innovation, in other words they have "how-to-knowledge" (Rogers, 1983: 167). Technological change is localised as consumers recombine their knowledge to adopt the technology. When the process of innovation is localised, the speed and direction of technology diffusion depends upon established knowledge (Metcalfe, 1981). At the consumer level, adoption is based on previous knowledge (Tonks, 1986). Thus, consumers familiar with a certain technology will quickly become familiar with the second generation innovation.

For the purposes of this study, we can assume that software voice can be included among the internet services. Since, the survey was conducted at the beginning of voice software service introduced in 2005 these adopters could be defined as earlier adopters within the Bass qualification. Indeed, the more technophile web-users are more willing to adopt this technology because they have the capabilities and are confident with the technology.

3. Data description

The data for the empirical econometric analysis were collected by the INSEE, the French National Institute for Statistics and Economic Studies in the 'Permanent survey on the life of households**,** information and communication technology (ICT)' in 2005. The survey comprises six sections: housing information, household characteristics, individuals' information, individuals' lifestyles, households' ICT equipment and individuals' ICT usage patterns. Some of the ICT related questions were included to provide EUROSTAT with information for monitoring ICT usage patterns in European countries. The data consist of a sample of 5,603 respondents[7] with the exception of the section related to personal characteristics which includes information from 13,410 individuals since this includes all household members aged 14 and over. This section was merged with the section on ICT usage controlling for IDENT_IND (individual identification). The data used for the empirical investigation relate to individuals' characteristics, household ICT equipment and ICT usage. These latter two sections are common across European countries.

The information used to constructed the dummy variable SOFTWAREVOICE, was based on responses to the question that asked individuals "Have you used the internet for calling during the last month (Skype, MSN)?". SoftwareVoice takes the value 1 if consumers use the internet for calling, 0 otherwise. The dummy variable TRIPLAY[8] captures individuals with home internet and takes the value 1 if households have broadband subscriptions with one of the two options triple play or double play, 0 otherwise. The initial database includes 5,603

[7] Since only one individuals per household has been interviewed.

[8] This variable was constructed from the HDEB variable created by INSEE and takes the value 1 if the household has a broadband connection and 0 if the household as a narrowband connection.

observations. Observations are dropped where the variable household income has missing values. Individuals without home internet connections and individuals who have never used the internet are also excluded as the questionnaire has constructed on the basis of filtering questions. The remaining sample corresponds to 1,745 observations. Table 3 presents the descriptive statistics for the variables. We have four set of explaining variable:

(i) Socio demographic variables. The socio-demographic variables identify the main characteristics of individuals and capture observed heterogeneity among individuals. The age of individuals is identified by four sets of dummies (AGE1, AGE2, AGE3, AGE4) taking as reference consumers over 51 years old. The variable capturing work categories is included with the dummy variables EMPLOYEE, SELF-EMPLOYED, STUDENT. The three dummy variables INC1500, INC2500, INC4000 indicate the households' revenue, the reference group are the households with more then 4,000 euros of revenues per month. The variable NBHOUSEHOLD takes the value 1 if individuals live with one or more people, 0 otherwise.

(ii) Geographical dwelling. The binary variables RURAL, URBAN100000, HBPLUS, PARIS capture the number of inhabitants, which could be a proxy for both information density and exchange (Goolsbee, Klenow, 2002) and a broadband accessible area at the time of the survey (the more populated areas had access to this service earlier). In higher population density areas the probabilities of being in contact with an adopters increase, hence to probabilities of getting information about the technology are higher.

(iii) Internet competences variables. The two binary variables COMPINT4 and COMPINT5 capture competences in e–related activities which could be associated with more technophile individuals. We dispose of other variables measuring the internet competences e.g. sending

email or searching information on the web, we choose to take in account only the variables giving information about more highly related competences.

(iv) Usage patterns. The set of variables capturing the usage patterns for voice communication are DAILY and WEEKLY which indicate respectively the volume of calls (in log) daily and weekly. Individuals indicated the number of calls daily, weekly and monthly (which is considered as the reference category).

Table 3: Description of the variables

DEPENDENT VARIABLES	DESCRIPTION	N	MEAN	MIN	MAX
Triplay	Equal to 1 if individuals have access to double or triple play	1745	0.320	0	1
Software Voice adoption	1 if individuals use internet for calling, 0 otherwise	1745	0.131	0	1
INDEPENDENT VARIABLES					
	Demographics (Rank effect variables)				
Gender (sexe)	Equal 1 if female, 0 otherwise	1745	0.489	0	1
Age1	Equal to 1 if age is between 15 to 20 , 0 otherwise	1745	0.116	0	1
Age2	Equal to 1 if age is between 21 to 30, 0 otherwise	1745	0.191	0	1
Age3	Equal to 1 if age is between 31 to 40, 0 otherwise	1745	0.273	0	1
Age4	Equal to 1 if age is between 41 to 50, 0 otherwise	1745	0.273	0	1
Employee	Equal to 1 if individual is employed in a firm , 0 otherwise	1745	0.407	0	1
Independent	Equal to 1 if individual is self employed , 0 otherwise	1745	0.064	0	1
Diploma (dipsupebis)	Equal to 1 if individual has a high school diploma, 0 otherwise	1745	0.087	0	1
Student	Equal to 1 if individual is a student , 0 otherwise	1745	0.151	0	1
Nbhousehold	Household number of 2 or more	1745	0.822	0	1
Inc1500	Equal to 1 if household's income<=1500 euros per month	1745	0.136	0	1
Inc2500	Equal to 1 if household's income<=2500 euros per month	1745	0.486	0	1
Inc4000	Equal to 1 if household's income<=4000 euros per month	1745	0.354	0	1
	Geographical location				
Rural	Equal to 1 if individual lives in countryside	1745	0.223	0	1
Urban100000	Equal to 1 if individual lives in a city with 100000 inhab. maximum	1745	0.149	0	1
Hdplus	Equal to 1 if individual lives in a city with more than 100000 inhab.	1745	0.290	0	1
Paris	Equal to 1 if individual lives in Paris	1745	0.220	0	1
	Internet related competencies				
Compint4	Equal to 1 if individual knows how to delete cookies and temporary files	1745	0.782	0	1
Compint5	Equal to 1 if individual knows how to create and modify a website	1745	0.245	0	1
	Calls volume				
Daily	Volume of calls make daily in log	1745	0.440	0	4.709
Weekly	Volume of calls make weekly in log	1745	0.581	0	4.382

4-The model and the results

Both the determinants of adoption and the substitution/complementarity effects are usually examined using the discrete choice setting. There are different approaches. (1) The nested model implies that different choices are bundled and the errors have a correlation dictated by the nested structure which has been applied into telecommunication studies by Train, McFadden, Ben-Akiva (1987). (2) The multivariate probit allows computing separate probit estimation which might have correlated disturbances, which suggest a substation effects among the two services (which we use). (3) The multiple discrete choice model enables to associate individuals with theirs demand according to the different characteristics of the choice made (Hendel, 1999) e.g. individuals with more internet competences are more willing to buy PC with particular characteristics.

Here, we use the bivariate probit as it allows to test the decision of adopting both services investigating on the pattern of substitution and complementarities. The bivariate probit is an extension of the univariate probit and it belongs among the class of multivariate models – option 2 (Maddala, 1983). The bivariate probit allows the two equations to have correlated disturbance leading to unobserved heterogeneity (Greene, 1998, 2002). This model applies the full information maximum likelihood estimation (FIML) (Jones, 2005) considering the joint distribution of the two variables. The general specification for simultaneous equation model is:

$$\begin{cases} y_1 = x_1\beta_{1n} + \varepsilon_1 \quad y_1 = 1 \;\; if \;\; y_1{}^* > 0, \qquad 0 \; otherwise \\ y_2 = x_2\beta_{2n} + \varepsilon_2 \quad y_2 = 1 \;\; if \;\; y_2{}^* > 0, \qquad 0 \; otherwise \end{cases}$$

$$E[\tilde{\varepsilon}_1 | x_1, x_2] = E[\tilde{\varepsilon}_2 | x_1, x_2] = 0$$

$$Var[\tilde{\varepsilon}_1 | x_1, x_2] = Var[\tilde{\varepsilon}_2 | x_1, x_2] = 1$$

$$Cov[\tilde{\varepsilon}_1, \tilde{\varepsilon}_2 | x_1, x_2] = \rho$$

Where y_1 and y_2 are vectors of the dependent variables, the latent dependent variable y_1^* et y_2^* are function of the utility that individuals gain from adopting either the two services or only one. The variables x_1 et x_2 are vectors of the independent variables and $\varepsilon_1, \varepsilon_2$ are vector of the unobserved effects.

Wilde (2000) shows that the repressors have to be exogenous and Monfardini and Radice (2008) demonstrate that the LR test is efficient for testing the exogenous nature of the variables[9]. We did the test for our variables. The explanatory variables could be the same for estimating the two equations, which is different from the probit with sample selection (Baum, 2006).

The bivariate probit has been already used in telecommunication studies. Greenstein (2000) used a trivariate and a bivariate probit to analyse the different strategies of internet providers in the US. Eisner and Waldon (2001) use a bivariate probit for analysing the joint decision of consumers to adopt both second line and online services. The bivariate probit analysis allows us to test the joint decision to adopt both services and at the same time it allows to evaluate the

[9] They refer essentially to the recursive model.

determinants influencing the adoption of each model. The structure of the variables frequency is shown in Table 4. Almost 8% of the individuals in our dataset chose to adopt both services.

The two independent variables are:

y_1 =1 if consumers subscribe to the Triplay service (TRIPLAY)

y_2 =1 if consumers use internet for calling (SOFTWAREVOICE).

The probability of each event occurs:

- subscribe to TRIPLAY services and use internet for calling $(y_{1i} = 1; y_{2i} = 1)$
- subscribe to TRIPLAY services and do not use internet for calling $(y_{1i} = 1; y_{2i} = 0)$
- do not subscribe to TRIPLAY services and use internet for calling $(y_{1i} = 0; y_{2i} = 1)$
- do not subscribe to TRIPLAY services and do not use internet for calling $(y_{1i} = 0; y_{2i} = 1)$

These probabilities are:

$$\Pr(y_1 = 1; y_2 = 1) = \Phi_2(x'_1\beta_1 + x'_2\beta_2, \rho)$$

$$\Pr(y_1 = 1; y_2 = 0) = \Phi_2(x'_1\beta_1, -x'_2\beta_2, -\rho)$$

$$\Pr(y_1 = 0; y_2 = 1) = \Phi_2[(-x'_1\beta_1, \beta_2 x'_2, -\rho)]$$

$$\Pr(y_1 = 0; y_2 = 0) = \Phi_2(-x'_1\beta_1, -x'_2\beta_2, \rho)$$

The Φ_2 stands for the standard bivariate normal cumulative distribution function (cdf). We use STATA 9 to compute our estimations with the command ‘biprobit’ which exploits the Newton – Raphson maximisation (Monfardini, Fabbri, 2007). The result of the bivariate probit

are interpreted in conjunction with the marginal effect reported in Table 7. The marginal effects for dummy variables are calculated on the basis of average probabilities[10].

The correlation coefficient ρ between the disturbances takes in account the existence of omitted variables or the unobserved heterogeneity (Savignac, 2008) which can influence simultaneously the adoption of SOFTWAREVOICE and the subscription to TRIPLAY. If $\rho \neq 0$, the two equations have to be estimated together. While, if $\rho = 0$ the errors are not correlated, thus the two equations should be analysed separately.

Table 4: Cross frequencies of the two independent variables

		Triplay		
		No subscription	*Subscription*	*Total*
Softwarevoice	*Do not usage*	1 096 *(62.81 %)*	420 *(24.07 %)*	1516
	Usage	91 *(5.21 %)*	138 *(7.91 %)*	229
		1 187	558	1 745

4.1 Presentation of the results

Table 6 presents the results of the complete model and underlines the interaction among different groups of variables. The signs and the significant *p*-values of variables are held stable as well as the LR test and the rho value. Table 5 presents the data per group in order to underline the effects of each set of variables. The comments refer to the estimation in Table 6 and 8.

[10] The software used STATA to identify the dummy variables and calculate impact effects.

The value of Rho (equal to 0.432) is positive which implies that there are common unobserved variables which positively influence the adoption of both SOFTWAREVOICE and TRIPLAY service. This confirms that individuals adopting both technologies have the propensity to use services enabling unlimited voice communications. In other words, the adopters of both technologies could be defined as innovators (as they are early adopters of these services). Greene (2002) indicates that the Wald test and the Lagrange test can be used to compute the independent test. Stata, the software we use, computes automatically the LR test testing for this independent test which is also an valuable test according to Monfardini and Radice (2008). The statistical Likelihood ratio test of independent equations, so called the LR test, rejects the null hypothesis that the two equations should be estimated separately. Since, the, the critical value computes on the chi-squared table is 3.84.

1st hypothesis

We hypothesise according to the literature (see section 2) that more technophile users have more probabilities to adopt SOFTWAREVOICE compared to the TRIPLAY adopters. As, we expected the COMPINT4 and COMPINT5 variables increase the probability to adopt the softvoice model but this does not have an impact on the adoption of the TRIPLAY which is consistent since subscribing to TRIPLAY does not require internet competences. This demonstrates that SOFTWAREVOICE application could be considered as an extension of other IP applications, in other words consumers customised with the IP related activities –first order innovation- have the ‘skill for using’ the second order innovation.

2nd hypothesis

Second, we test the substitution and complementarity effects among the two services offering voice services to reduced costs. The result of the bivariate probit demonstrates they have different conditions of adoption and different usage patterns. It emerges that they are imperfect substitutes. Indeed, both allow voice communication.

As demonstrated on table 5 (column (a)) and on the table 6, individuals belong to the class of AGE1 and AGE2 have more probabilities to adopt SOFTWAREVOICE. While, individuals belong to the class of AGE2 and AGE3 have more probabilities to live in households having subscription to TRIPLAY services. In other words younger people have probabilities to adopt both services. There is not effect of revenue variables as the adopt of these two services have more usages patterns influence then revenues, different might be the case for the adoption of the mobile phone where the revenues can have an important impact.

Being self-employed increases the probability of adopting SOFTWAREVOICE, might be this service is a good work tool. The variables measuring the incomes do not seem to have an influence on adoption decision. On the other hand, the diffusion of services such as SOFTWAREVOICE leads to more unobserved heterogeneity, as capturing the propensity to use this technology might lead to ability or propensity to use e-technology. To live with one or more individuals MORE significantly increases the probability to subscribe to the triple play offer as the household will profit from unlimited access to voice communication. The volumes of calls DAILY have positive and significant effects on the probability of subscribing to triple or double play option, which might justify the decision of subscribe to this services. We do not have information about the particular preferences of consumers toward this bundled offer.

We estimate also a recursive bivariate probit model for determining the substitution/complementarities effect on population treated. But the recursive bivariate model can not be efficiently computed. Since, the recursive bivariate model implies that two dependent variables should have causality effect (Maddala, 1983, Monfardini, Fabbri, 2007) and the treated population sample is quite small.

3rd hypothesis

The variables capturing urban density enter through four dummy variables (Urban20000 is the omitted category- city with less then 20000 inhabitants). As expected COUNTRYSIDE has significant and negative effect on subscribing to TRIPLAY because this service was not widely available in rural areas, while, Urbanplus and PARIS are positive and significant. Since, at the time of the survey this service was available in most populated area. It also is a major of network effect as individuals might want to joint the network where relatives are. The magnitude effects of living in these areas are respectively 6% and 9.4%.

The variables capturing urban concentration can be interpreted as a source of both epidemic and network effects. In the former case, the variables can be considered as a source of the geographical proximity indicating the density of information exchange among individuals. When innovation adoption is driven by network effects, the adoption decision is driven by the necessity to be connected with other and utility of joining a network is associated to the numbers of individuals presented into the network. Unfortunately, the survey gives no information on either sources of information about the technology or the typology of network effects supporting innovation diffusion.

In terms of policy, increased diffusion of internet connection and broadband access could reduce the digital divide among regions, which might reduce the effect of state dependence considered as a source of serial persistence (Demoussis, Giannakopoulos, 2006). This implies that individuals who choose to not subscribe to TRIPLAY make they are not used to it. They might decide to subscribe if the area dwelling has provided with the TRIPLAY services, or if they receive a subvention to adopt it. On the contrary, when the non adoption leads to unobserved heterogeneity this might be related to personal characteristics e.g. refusal to use technology, no internet ability, which implies that policy actions can hardly affect the decision. This can be the case for non adopters of SOFTWAREVOICE service.

Table 5: Estimation of the Bivariate Probit

	(a)	(b)	(c)	(d)	(e)	(f)
Softwarevoice (constant)	-1.283	-1.188	-1.054	-1.121	-1.516	-1.140
Gender	-	-	-	-	-	-
Age1	0.376 (0120)***	-	-	-	-	-
Age2	0.320 (0103)***	-	-	-	-	-
Age3	0.170 (0.096)*	-	-	-	-	-
Age4	0.560 (0.126)	-	-	-	-	-
Student	-	0.267 (0.107)**	-	-	-	-
Self-employed	-	0.362 (0.146)**	-	-	-	-
Employee	-	-0.004 (0.085)	-	-	-	-
Diploma	-	-0.008 (0.078)				
More	-	-	-0.046 (0.107)	-	-	-
Inc1500	-	-	0.144 (0.127)	-	-	-
Inc2500	-	-	-0.112 (0.085)	-	-	-
Inc4000	-	-	-0.019 (0.085)	-	-	-
Countryside	-	-	-	-0.216 (0.132)	-	-
Urban100000	-	-	-	-0.086 (0152)	-	-
Urbanplus	-	-	-	0.075 (0.119)	-	-
Paris	-	-	-	0.122 (0.124)	-	-
Compint4	-	-	-	-	0.365 (0.110)***	-
Compint5	-	-	-	-	0.329 (0.085)***	-
Daily	-	-	-	-	-	0.033 (0.054)
Weekly	-	-	-	-	-	0.008 (0.044)
Triplay (constant)	-0.615	-0.474	-0.645	-0.527	-0.457	
Gender	-	-	-	-	-	-
Age1	0.183 (0.103)*	-	-	-	-	-
Age2	0.321 (0.085)***	-	-	-	-	-
Age3	0.220 (0.077)***	-	-	-	-	-
Age4	-0.040 (0.099)	-	-	-	-	-
Student	-	0.017 (0.093)	-	-	-	-
Self-employed	-	-0.081 (0.133)	-	-	-	-
Employee	-	-0.008 (0.068)	-	-	-	-
Diploma		-0.025 (0.064)	-	-		-
More	-	-	0.193 (0.091)**	-	-	-
Inc1500	-	-	0.088 (0.110)	-	-	-
Inc2500	-	-	0.056 (0.069)	-	-	-
Inc4000	-	-	0.051(0.069)	-	-	-
Countryside	-	-	-	-0.324 (0.109)***	-	-
Urban100000	-	-	-	-0.129 (0.126)	-	-
Urbanplus	-	-	-	0.199 (0.099)**	-	-
Paris	-	-	-	0.323 (0.104)****	-	-
Compint4	-	-	-	-	-0.070 (0.078)	-
Compint5	-	-	-	-	0.179 (0.074)**	-
Daily	-	-	-	-	-	0.126 (0.048)***
Weekly	-	-	-	-	-	0.032 (0.036)
LR test rho=0 chi2(1)	87.1752	92.2197	91.1233	84.9299	88.4002	89.0472
RHO	0.432	0.444	0.441	0.426	0.440	0.436

standard error *(.)* p<10 %(*), p<5 %(**), p<0.1 % (***)

Table 6: Estimation of the Bivariate Probit

	(1)	(2)	(3)	(4)	(5)	(6)
Softwarevoice (constant)	-1.521	-1.200	-1.205	-1.073	-1.106	-1.479
Gender	-0.168 (0.081)**	- 0.247(0.077)***	-0.242 (0.078)***	-0.231 (0.078)***	-0.238 (0.079)***	-0.164 (0.081)**
Age1	0.338 (0.207)	0.405 (0.135)***	0.347 (0.201)*	0.362 (0.202)*	0.364 (0.204)*	0.322 (0.206)
Age2	0.366 (0.133)***	0.351 (0.120)***	0.366 (0.127)***	0.383 (0.130)***	0.366 (0.131)***	0.350 (0.132)***
Age3	0.223 (0.123)*	0.198 (0.115)*	0.221 (0.119)*	0.241 (0.121)**	0.232 (0.121)*	0.212 (0.123)*
Age4	0.034 (0.134)	0.055 (0.126)	0.040 (0.132)	0.045 (0.132)	0.034 (0.133)	0.028 (0.134)
Student	0.019 (0.167)	-	0.075 (0.162)	0.056 (0.163)	0.030 (0.165)	0.012 (0.167)
Self-employed	0.359(0.156)**	-	0.329 (0.151)**	0.316 (0.151)**	0.372 (0.153)**	0.363 (0.156)**
Employee	-0.078 (0.093)	-	-0.079 (0.090)	-0.081 (0.091)	-0.084 (0.091)	-0.084 (0.093)
Diploma	-0.117 (0.088)	-	-0.006 (0.082)	-0.039 (0.086)	-0.073 (0.087)	-0.115 (0.088)
More	-0.032 (0.115)	-	-	-0.081 (0.113)	-0.043 (0.114)	-0.027 (0.116)
Inc1500	0.015 (0.140)	-	-	0.032 (0.137)	0.059 (0.138)	0.020 (0.140)
Inc2500	-0.131 (0.092)	-	-	-0.155 (0.090)*	-0.132 (0.091)	-0.129 (0.092)
Inc4000	-0.002 (0.089)	-	-	-0.024 (0.087)	- 0.001 (0.088)	-0.003 (0.087)
Countryside	-0.181 (0.136)	-	-	-	-0.210 (0.134)	-0.177 (0.135)
Urban100000	-0.075 (0.156)	-	-	-	-0.090 (0155)	-0.071 (0.156)
Urbanplus	0.091 (0.122)	-	-	-	0.081 (0.121)	0.092 (0.122)
Paris	0.155 (0.131)	-	-	-	0.151 (0.129)	0.164 (0.130)
Compint4	0.345 (0.113)***	-	-	-	-	0.345 (0.113)***
Compint5	0.238 (0.089)***	-	-	-	-	0.241 (0.089)***
Daily	0.060 (0.058)	-	-	-	-	-
Weekly	0.028 (0.045)	-	-	-	-	-
Triplay (constant)	-0.913	-0.593	-0.571	-0.707	-0.838	-0.816
Gender	-0.028 (0.066)	-0.005 (0.063)	-0.011 (0.063)	-0.016 (0.064)	-0.029 (0.064)	-0.017 (0.066)
Age1	0.276 (0.176)	0.165(0.113)	0.279 (0.170)	0.257 (0.171)	0.261 (0.175)	0.239 (0.175)
Age2	0.355 (0.108)***	0.303 (0.097)***	0.340 (0.103)***	0.339 (0.105)***	0.315 (0.106)***	0.316 (0.107)***
Age3	0.208 (0.097)**	0.201 (0.090)**	0.222 (0.094)**	0.206 (0.095)**	0.191 (0.096)**	0.184 (0.097)*
Age4	-0.037 (0.105)	-0.040 (0.099)	-0.017 (0.102)	-0.029 (0.103)	-0.054 (0.104)	-0.053 (0.104)
Student	-0.183 (0.145)	-	- 0.144 (0.140)	-0.147 (0.141)	-0.196 (0.144)	-0.197 (0.144)
Self-employed	0.015 (0.139)	-	-0.069 (0.137)	-0.073 (0.137)	0.032 (0.139)	0.025 (0.139)
Employee	-0.038 (0.074)	-	-0.056 (0.072)	-0.051 (0.073)	-0.062 (0.074)	-0.054 (0.074)
Diploma	-0.066 (0.072)	-	-0.003 (0.067)	0.011 (0.070)	-0.054 (0.071)	-0.060 (0.072)
More	0.254 (0.097)***	-	-	0.184 (0.095)*	0.266 (0.096)***	0.264 (0.096)
Inc1500	0.061 (0.118)	-	-	0.026 (0.116)	0.080 (0.118)	0.069 (0.118)
Inc2500	0.057 (0.074)	-	-	0.018 (0.0702)	0.065 (0.074)	0.062 (0.074)
Inc4000	-0.032 (0.072)	-	-	-0.078(0.070)	-0.034 (0.071)	-0.034 (0.071)
Countryside	-0.321 (0.111)***	-	-	-	-0.317 (0.110)***	-0.310 (0.111)***
Urban100000	-0.138 (0.128)	-	-	-	-0.130 (0.128)	-0.130 (0.128)
Urbanplus	0.220 (0.101) **	-	-	-	0.226 (0.100)**	0.222 (0.100)**
Paris	0.355 (0.107)***	-	-	-	0.377 (0.106)***	0.376 (0.106)***
Compint4	-0.070 (0.081)	-	-	-	-	-0.070 (0.081)
Compint5	0.138 (0.078)*	-	-	-	-	0.145 (0.078)*
Daily	0.140 (0.048)***	-	-	-	-	-
Weekly	0.057 (0.037)	-	-	-	-	-
LR test rho=0 chi2(1)	81.976	88.9216	88.9352	89.9164	84.384	83.3164
RHO	0.432	0.439	0.439	0.443	0.431	0.434

standard error *(.)* p<10 %(*), p<5 %(**), p<0.1 % (***)

Table 7: Marginal effects

	Mfx (11)	Mfx (10)	Mfx (01)
Gender	-0.018	-0.015	0.008
Age1	0.058	0.019	0.044
Age2	0.065	0.017	0.066
Age3	-0.035	0.011	0.039
Age4	0.001	0.005	-0.015
Student	-0.007	0.011	-0.055
Self-employed	0.041	0.043	-0.036
Employee	-0.010	-0.006	-0.004
Diploma	-0.015	-0.008	-0.008
More	0.010	-0.017	0.076
Inc1500	0.004	-0.002	0.017
Inc2500	-0.010	-0.015	0.031
Inc4000	-0.002	0.001	-0.009
Countryside	-0.031	-0.003	-0.077
Urban100000	-0.014	-0.001	-0.034
Urbanplus	0.021	-0.002	0.058
Paris	0.036	-0.004	0.094
Compint4	0.029	0.031	-0.055
Compint5	0.034	0.017	0.0159
Daily	0.013	-0.001	0.036
Weekly	0.006	-0.0001	0.014

5. Probit specification for the software voice application

To examine the usage of the SOFTWAREVOICE implies a deeply investigation of the adoption pattern. Hence, we use a univariate probit model for testing the effect of experience in internet usage.

$$\text{prob[softwarevoice=1]} = \Phi(\alpha_0 + \alpha_1 \text{gender}_i + \alpha_2 \text{age1}_i + \alpha_3 \text{age2}_i + \alpha_4 \text{aged3}_i + \alpha_5 \text{aged4}_i + \alpha_6 \text{student}_i + \alpha_7 \text{selfemployed}_i + \alpha_8 \text{employee} + \alpha_9 \text{diploma} + \alpha_{10} \text{nbhousehold}_i + \alpha_{11} \text{inc1500}_i + \alpha_{12} \text{inc2500}_i + \alpha_{13} \text{inc4000}_i + \alpha_{14} \text{countryside}_i + \alpha_{15} \text{urban100000}_i + \alpha_{16} \text{urbanplus}_i + \alpha_{17} \text{paris}_i + \alpha_{18} \text{compint4}_i + \alpha_{19} \text{compint5}_i + \alpha_{20} \text{debnet1}_i + \alpha_{21} \text{debnet2}_i + \alpha_{22} \text{debnet4}_i + \alpha_{23} \text{debnet5}_i + \alpha_{24} \text{daily}_i + \alpha_{25} \text{weekly}_i)$$

The dummy variables Debnet5 takes value if individuals use internet for 10 year, debnet4 has a value 1 if individual has experience into internet going from 10 to 5 years. We use as reference group the individuals having experience into internet between 3 and 5 years (debnet3). The results are showed in table 8.

Table 8. Probit estimation for testing the experience on internet usages

	(1)	(2)
Gender	-0.210**	-0.210**
	(0.075)	(0.073)
Age1	0.287	0.320+
	(0.186)	(0.181)
Age2	0.368**	0.364**
	(0.125)	(0.121)
Age3	0.248*	0.254*
	(0.116)	(0.114)
Age4	0.101	0.132
	(0.128)	(0.125)
Student	0.162	0.164
	(0.145)	(0.140)
Self-employed	0.362*	0.372*
	(0.149)	(0.145)
Employed	-0.137	-0.135
	(0.085)	(0.084)
Diploma	-0.065	-0.022
	(0.084)	(0.081)
Nbhousehold	0.051	0.071
	(0.104)	(0.090)
Inc1500	-0.039	-
	(0.129)	
Inc2500	-0.138	-
	(0.089)	
Inc4000	0.011	-
	(0.085)	
Countryside	-0.129	-0.108
	(0.127)	(0.126)
Urbain100000	-0.059	-0.046
	(0.146)	(0.144)
Hbplus	0.109	0.157
	(0.114)	(0.112)
Paris	0.195	0.240*
	(0.124)	(0.121)
Compint4	0.334**	0.312**
	(0.103)	(0.100)
Compint5	0.254**	0.243**
	(0.084)	(0.083)
debnet1	0.212	0.199
	(0.136)	(0.134)
debnet2	0.056	0.084
	(0.103)	(0.100)
debnet4	0.150	0.172*
	(0.096)	(0.095)
debnet5	0.412**	0.445**
	(0.154)	(0.152)
Daily	0.111*	0.100*
	(0.054)	(0.053)
Weekly	0.044	0.033
	(0.043)	(0.042)
_cons	-1.888**	-2.012**
	(0.212)	(0.188)
N	2377	2462
Degree of freedom	25	22
LR test	108.10	109.56
Pseudo R2	0.0678	0.0665

Standard errors in parentheses
+ p<.10, * p<.05, ** p<.01

Using internet for more then 10 years positively influences the likelihood to adopt SOFTWAREVOICE, confirming that users of first generation innovations have more to adopt further applications. These individuals were the first internet users 'early adopters'. Into the usage of this technology experience on the internet does not have great impact compared to internet related competences. This confirms that users of the first order innovation are customised with latter innovations. What can determine the cut off point among the different group of users? The first internet adopters are more customised with the internet incremental innovations and they have confidence into the internet application. As mentioned previously in 2005, the software voice application has not yet been developed, thus the adopters are the innovators or the early adopters. The estimation of the column (1) has less observation as there are some missing data for the income variables.

Into the specification into column (2) where all sample is considered, the geographic density in other word individuals living in Paris have more probabilities to adopt SOFTWARE VOICE which we interpret as a source of local information exchange.

6. Limitations

One of the main limitations of our study is related to the lack of information about the needs driving the adoption of SOFTWAREVOICE. We do not have information neither about the language spoken or the frequency of travel which might give to us more information about the characteristics of diffusion pattern of this technology which could break geographical distance as it has worldwide diffusion. On the other hand, the survey does not give information on telecommunication operators chosen by households' e.g. new entrants or incumbents. This will be extremely valuable for analysing the churn consumers' propensity.

There were no questions referring to preferences for bundled services offered by the triple play providers. Here, we suppose that the individuals subscribing to the TRIPLAY, because

of the earlier development of this option by France Telecom. We can hypothesis that individuals subscribing to TRIPLAY have chosen new entrants, thus they could be defined as ‘early adopters’ (Rogers, 1983; Dickerson, Gentry, 1983). They might churn from incumbents to new entrants. If the survey would contain all these information, it can be more useful for the telecommunication policy authorities and for the Ministry of Research and Public Administration. In addition, the cross section nature of the sample could not give us more information about the dynamics of the adoption.

Conclusion

The telecommunication industry has seen the emergence of numerous new technologies, including VoIP which permits unlimited voice traffic. From the empirical findings, it emerges that the two services, namely Softwarevoice and Triplay services are competing for the same group of consumers in other words experienced internet users and youngest people, which might be defined as ‘early innovators’ (Rogers, 1983). The two services could be considered as imperfect substitutes. The imperfect substitute feature is the consequences of the consumers’ perceptions and differentiation among the two models. It is not possible to determine if one of the two patterns can dominate the market. This opens up the possibility of persistence plurality of services, such as the case of Apple in the PC industry (Swann, 2002).

On the one hand, both services enable voice traffics completely free or at reduced tariffs. On the other hand, the Software service applications require that individuals should have some IP related competences and thus confidence with using the Internet. At the same time, it enables individuals to have access to other communication services such as video conferencing, instant messaging and so forth. Meanwhile, the adoption of TRIPLAY is highly geographical determined, as it is high density areas that have access to this service. The results give insights also on the debate on the digital divide, as at this time the rural area were

discriminate on the access to ICT and broadband connection. As concerned the network effect pattern, the adoption of TRIPLAY is mainly driven by general network effects whilst SOFTWAREVOICE adoption can be motivated by social network effects. The article aims to add further questions at the policy and regulatory perspective:

(1) Effect of substitutability and complementarity. Both the two services enable free communication or at reduced tariffs fostering the displacement of the traditional bill and keep tariffs. From the emergence of the technology, the policy debate has been concentrated on the effect of substitutability and complementarity among the Triplay and the traditional PSNT network. The French national agency of regulation (and generally in Europe) does not include software voice among the telecommunication services but among the information one. However, some institutions have forbidden the use of SOFTWAREVOICE service, which open up questions relating to the net neutrality.

(2) The diffusion of broadband as mentioned on the comments can foster the diffusion of the TRIPLAY service giving access to larger number of individual to unlimited voice services. Here, the diffusion of the services is more related to general network effects. While, the adoption of SOFTWAREVOICE service leads more to unobserved individual characteristics, because it requires to individuals to have competences on internet applications and confidence on internet as a tool of communication exchange. Moreover, the existence of social network effects is extremely important for the diffusion of this internet service.

(3) The diffusion of alternative broadband technology such as the WIMAX can completely change the telecommunication sector and voice communication. The Softwarevoice can become easily accessible to anyone. The diffusion of the

wireless technology might oblige the authority to regulate this service in order to safeguard the consumers. The incumbents will be foster finally to change their business model giving free access to free voice traffics.

Appendix A: Descriptive statistics

Independent variables	Description	Software=1 (229 obs.)	TRIPLAY=1 (558 obs.)
Gender (sexe)	Equal 1 if female	0.603 (0.490)	0.489 (0.500)
Age1	Equal to 1 if age is between 15 to 20 , 0 otherwise	0.393 (0.489)	0.120 (0.325)
Age2	Equal to 1 if age is between 21 to 30, 0 otherwise	0.157 (0.365)	0.229 (0.421)
Age3	Equal to 1 if age is between 31 to 40, 0 otherwise	0.240 (0.428)	0.296 (0.457)
Age4	Equal to 1 if age is between 41 to 50, 0 otherwise	0.279 (0.550)	0.165 (0.371)
Employee	Equal to 1 if individual is employed in a firm , 0 otherwise	0.367 (0.483)	0.407 (0.492)
Independent	Equal to 1 if individual is self employed , 0 otherwise	0.100 (0.301)	0.059 (0.236)
Diploma	Equal to 1 if individual has a high school diploma, 0 otherwise	0.445 (0.498)	0.462 (0.499)
Student	Equal to 1 if individual is a student , 0 otherwise	0.205 (0.405)	0.154 (0.361)
Nbhousehold	Household number of 2 or more	0.799 (0.401)	0.846 (0.361)
Inc1500	Equal to 1 if household's income<=1500 euros per month	0.183 (0.388)	0.136 (0.343)
Inc2500	Equal to 1 if household's income<=2500 euros per month	0.345 (0.476)	0.412 (0.493)
Inc4000	Equal to 1 if household's income<=4000 euros per month	0.332 (0.472)	0.344 (0.475)
Rural	Equal to 1 if individual lives in countryside	0.161 (0.369)	0.140 (0.347)
Urban100000	Equal to 1 if individual lives in a city with 100000 inhab. maximum	0.096 (0.295)	0.090 (0.286)
Hdplus	Equal to 1 if individual lives in a city with more than 100000 inhab.	0.332 (0.472)	0.342 (0.475)
Paris	Equal to 1 if individual lives in Paris	0.262 (0.441)	0.288 (0.453)
Daily	Volume of calls make daily in log	0.471 (0.819)	0.509 (0.812)
Weekly	Volume of calls make weekly in log	0.574 (0.909)	0.574 (0.954)

Standard deviation in ()

Bibliography

Antonelli, C., 2006. Diffusion as a process of creative adoption, Journal of Technological Transfer, 31, 211-226.

Andreosso, B., Jacobson, D., 2005. Industrials economics & organization a European Perspective. McGraw-Hill Education, Second Edition.

Arthur, B., 1989. Competing technologies increasing returns and lock-in by small historical events. Economic Journal 99, 116-131.

Asterbro, T., 2002. Noncapital investment costs and the Adoption of CAD and CNC in U.S. metalworking industries. The RAND Journal of Economics 33 (4), 672-688.

Bass, F. M., 1969. A new product growth model for consumer durables. Management Science 15 (January), 215-227.

Bass F. M., 1980. The Relationship between diffusion rates, experience curves, and demand elasticities for consumer durable technological innovations. Journal of Business 53 (July-part 2), 51-67.

Battisti, G., Stoneman P., 2005. The intra-firm diffusion of new process technologies. International Journal of Industrial Organisation 23, 1-22.

Baum, C., 2006. Introduction to Econometrics Using Stata, Stata Press.

Birke, D., Swann, G. M. P., 2006. Network Effects and the choice of mobile phone operator. Journal of Evolutionary Economics (16), 65-84.

Bochet, R., Brossar, O., 2007. The Variety of ICT adopters in the intra-firm diffusion process: theoretical arguments and empirical evidence. Structural Change and Economic Dynamics 18, 409-437.

Bresnahan, T. F., Trajtenberg, M., 1995. General Purpose Technologies 'Engines of growth'?. Journal of Econometrics 65, 83-108.

Church, J., Ware, R. 1998. Network Industries, Intellectual Property Rights and competition policy, in R.D. Anderson and N. T. Gallini (Eds.), Competition Policy and Intellectual Property Rights in the Knowledge-Based Economy, Canada, University of Calgary

Constantiou, D. I., Kautz, K., 2008. Economics factors and diffusion of IP telephony: Empirical evidence from an advanced market. Telecommunication Policy 32 (3-4), 197-211.

Corrocher, N., 2003. The diffusion of Internet telephony among consumers and firms: current issues and future prospects. Technological Forecasting and Social Change 70 (6), 525– 544.

David, P.A., 1985. Clio and the economics of QWERTY. American Economic Review 75, 332-337.

Demoussis, M., Giannakopoulos, N., 2006. The dynamics of computer ownership in Greece. Information Economics and Policy 18, 73-86.

Dickerson, M. D., Gentry, J. W., 1983. Characteristics of adopters and non adopters of home computers. Journal of Consumer Research 10, 225-235.

Duffy-Deno, K. T., 2001. Demand for additional telephone lines: an empirical note. Information Economics and Policy 13, 283-299.

Economides, N., 1995. Network Externalities, Complementarities, and Invitations to Enter. http://www.stern.nyu.edu/networks/ejpe95.pdf

Eisner J., Waldon T., 2001. The demand for bandwidth: second telephone lines and on-line services. Information and economics policy 13, 301 -309.

Faria, A., Fenn, P., Bruce, A., 2002. Determinants of adoption of flexible production technologies: evidence from Portuguese manufacturing industry. Economics of Innovation and New Technology 11 (6), 569-580.

Fildes, R., Kumar, V., 2002. Telecommunications demand forecasting -a review. International Journal of Forecasting 18, 489-522.

Fransman, M., 2003. Evolution of the Telecommunications Industry. In: Madden, G, (eds.) The International Handbook of Telecommunications Economics. Aldershot: Edward Elgar, Vol. III, pp. 15-3.

Garbacz, C., Thompson, H. G., 2007. Demand for telecommunication services in developing countries. Telecommunication policy 31, 276-289.

Geroski, P. A., 2000. Models of technology diffusion. Research Policy 29, 603-625.

Gentzkow, M., 2007. Valuing new goods in a model with complementarity: online newspapers. American Economic Review 97 (3), 713-744.

Goldfard A. , Prince J., 2008. Internet adoption and usage patterns are different: implications for digital divide. Information and economics and policy 20, 2-15.

Goolsbee, A., Klenow, P.J, 2002. Evidence on learning and network externalities in the diffusion of home computers, Journal of Law and Economics 45, 317-342.

Greeinstein, S., Mazzeo, M., 2006. The role of differentiation strategy in local telecommunication entry and market evolution: 1999-2002. The Journal of Industrial Economics 54 (3), 323-350.

Greeinstein, S., 2000. Building and delivering the virtual world: commercializing services for internet access. The journal of Industrial economics 48 (4), 391-411.

Greene, W. H., 2002. Econometric Analysis, Prentice-Hall International.

Greene, W.H., 1998. Gender economics courses in liberal arts colleges: further results. Journal of economic education 29 (4), 291-300.

Hamilton, J., 2003. Are main lines and mobile phones substitutes or complements? Evidence from Africa. Telecommunication Policy 27, 109-133.

Hendel, 1999. Estimating Multiple-Discrete Choice Models: An Application to Computerization Returns. Review of Economic Studies 66, 423-446.

Hong, S. H., 2007. The recent growth of the internet and changes in household-level demand for entertainment. Information Economics and Policy 19, 304-318.

Jones, M. A., 2005. Applied econometrics for health economists. Office of Health Economist, Second edition.

Karshenas, M., Stoneman, P. L., 1993. Rank, Stock, Order, and Epidemic Effects in the Diffusion of New Process Technologies: An Empirical Model. The RAND Journal of Economics 24 (4), 503-528.

Katz, L. M., Shapiro, C., 1986. Technology Adoption in the Presence of Network Externalities. The Journal of Political Economy 94 (4), 822-841.

Klemperer, P., 1995. Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade. The Review of Economic Studies 62(4), 515–539.

Klepper, S. Graddy, E., 1990, The evolution of new industries and the determinants market structure. RAND Journal of Economics 21 (1), 27-44.

Krafft, J., 2004. Knowledge intensive life cycles: the case of telecommunications. ISS Conference Milan 9-12 June.

Krafft, J., 2003. Vertical structure of the industry and competition an analysis of the evolution of the info-communication industry. Telecommunication Policy 27 (8-9), 625-649.

Le Guel, F., Pennard, T., Suire, R., 2005. Adoption et usage marchand sur Internet: une etude économqetrique sur données bretonnes. Economie et prévision 167, 67-84.

Maddala, G. S., 1983. Limited dependent and qualitative variables in econometric, Cambridge University Press.

Majumdar, S. K, Venkataraman, S., 1998. Network effects and the adoption of new technology: evidence from the U.S. telecommunication industry. Strategic Management Journal 19, 1045-1062.

Mansfield, E., 1961. Intra-firm Rates of Diffusion of an Innovation. The Review of Economics and Statistics 45 (4), 348-359.

Metcalfe, J. S., 1981. Impulse and diffusion in the study of technical change. Futures (B) 347-359.

Michalakelis, C., Varoutas, D., Sphicopoulos, T., 2008. Diffusion models of mobile telephony in Greece. Telecommunication Policy 32 (3-4), 234-245.

Mindel, J. L., Sicker, D. C., 2006. Leveraging the EU regulatory framework to improve a layered policy model for US telecommunications markets. Telecommunication Policy 30 (2), 136-148.

Monfardini, C., Radice, R., 2008. Testing exogeneity in the bivariate probit model: a Monte Carlo study. Oxford bulletin of economics and statistics 70 (2), 271-282.

Monfardini, C., Fabbri,C., 2007. Style of practice and assortative mating: a recursive probit analysis of Caesarean section scheduling in Italy. Applied Economics (1), 1-13.

Mowery, D.C., T.S., Simcoe, 2002. Is the Internet a U.S. Invention? – An Economic and Technological History of Computer Networking. Research Policy 31 (8-9), 1369-1387.

Pleatsikas, C., Teece, D., 2001. The analysis of market definition and market power in the context of rapid innovation. International Journal of Industrial Organization 19, 665-693.

Robertson, A., Soopramanien, D., Filde, R., 2007. Segmental new-product of residential broadband services. Telecommunication Policy 31, 265-275.

Rodini, M., Ward, M. R., Woroch, G. A., 2003. Going mobile: substitutability between fixed and mobile access. Telecommunication Policy 27, 457-476.

Rogers, E. M., 1983. Diffusion of Innovations. New York: The free Press, The third edition.

Rohlfs, J., 1974. A Theory of Interdependent Demand for a Communications Service. The Bell Journal of Economics and Management Science 5(1), 16-37.

Roux, P., Galliano, D., 2008. Organisational Motives and Spatial Effects in Internet Adoption and Intensity of Use: Evidence from French Industrial Firms. The Annals of Regional Science 42 (2), 425-448.

Savignac, F., 2008. The impact of financial constraints on innovation: what can be learned from a direct measure?. Economics of Innovation and New Technology, forthcoming.

Shapiro, C., Varian, H. R., 1998. Information rules. Harvard Business School Press, Boston.

Steinmueller, E. W., 2006. Learning in the knowledge-based economy: the future as viewed from the past. in: Antonelli, C., Foray, D., Hall, H.B., Steinmueller, W. E., (Eds.) New frontiers in the economics of innovation and new technology. Edward Elgar, Nothampton, MA, USA

Steyer, A., Zimmermann, J. B., 2004. Influence sociale et diffusion de l'innovation. Mathématiques et sciences humaines 168, special edition Les réseaux sociaux.

Swann, P. M. G., 2002. The functional form of network effects. Information Economics and Policy 14 (3), 417-429.

Train, K. E., McFadden, L. D., Ben-Akiva, M., 1987. The demand for local telephone service: a fully discrete model of residential calling patterns and service choices. RAND Journal of Economics 18 (1), 109-123.

Tonks, I., 1986. The Demand for Information and the Diffusion of new Product. International Journal of Industrial Organization 4, 397-408.

Torre, A., Rallet, A., 2005. Proximity and Localization. Regional Studies 39 (1), 47-59.

Van den Ende, J., Dolfsma, W., 2005. Technology-push, demand-pull and the shaping of the technological paradigms-Patterns in the development of Computing Technology. Journal of Evolutionary Economics 15, 83-99.

Wilde, J., 2000. Identification of multiple equation probit models with endogenous dummy regressors. Economics Letters 69, 309-312.

The Patent Quality Control Process: Can We Afford An (Rationally) Ignorant Patent Office?

Jing-Yuan Chiou*

February 2008

Abstract

This paper considers patent granting as a two-tiered process, which consists of patent office examination and court challenges. It argues that, when the patent-holder has private information about the patent validity, (i) a weak patent is more likely to be settled and thus escape court challenges than a strong patent; and (ii) a tighter examination by the patent office may strengthen private scrutiny over a weak patent. Both work against Lemley (2001)'s hypothesis of a "rationally ignorant" patent office. The paper also considers application fees and a pre-grant challenge procedure, and shows that the former, used as a tool to deter opportunistic patenting, may crowd out private enforcement but cannot replace public enforcement; while the usefulness of the latter is subject to several restrictions, including the private challenger's timing choice.

Keywords: Case Selection, Patent Quality, Public and Private Enforcement of Law.
JEL codes: K40, O31, O34

*Assistant Professor, IMT Lucca. This is a substantive revision of the second chapter of my Ph.D. dissertation submitted to the Universitè de Toulouse 1. I would like to thank Bernard Caillaud, Vincenzo Denicolò, and Jean Tirole for very helpful comments. Special thanks go to Jean Tirole for his continuing support and encouragement. All errors are mine. Comments are welcome, please send to: jy.chiou@imtlucca.it

1 Introduction

The patent granting process is often described as a two-tiered system: Besides the inspection by patent office examiners (the public enforcement tier), private parties can also challenge the validity of issued patents in court or at the patent office (the private enforcement tier).[1] Indeed, private challengers are usually thought to have significant advantages over the public agency. They have more knowledge about which patents cover valuable inventions, so the granted monopoly entails serious consequences; they also closely follow technological developments and have more information about where to locate those prior arts useful in making patent granting decisions. Reflecting upon this view, Lemley (2001) advocates a "rationally ignorant" patent office, and argues that instead of carefully scrutinizing every patent application at the patent office, it would be more efficient to lower the examination standard and issue some patents with questionable quality, while letting private parties select which patents to dispute in court. A glance at the United States Patent Reform Act of 2007 also reveals this emphasis on the private sector to eliminate weak patents.

In this paper, we argue that this "rational ignorance" hypothesis ignores both private players' strategic behavior and how public efforts would affect private enforcement. Despite the advantages, private parties frequently settle cases, leaving the contested patents in force. Among those unsettled cases, the disputed patents may be systematically biased toward certain characteristics. This "case selection," as we will show in this paper, constrains the effectiveness of private force and needs to be taken into account in order to induce proper cooperation between private and public sectors in the patent quality control process.

We consider a situation where, before launching a validity challenge, the settlement bargaining between the patent-holder and a potential challenger is clouded with asymmetric information. That is, the patent-holder has some private information about the validity of the disputed patent. We use a simple two-type model where the patent-holder has either a strong or a weak patent (section 2), and the challenger optimally chooses his litigation efforts if bargaining breaks down. Fixing the litigation effort, a strong patent, assumed to be possessed by a true inventor, is more likely to withstand challenges. By contrast, a weak patent is more likely to be invalidated in court because, as an opportunistic player, its owner tried to patent an already existing technology.

[1] To challenge at the patent office, a private party can request patent reexamination in the United States, and patent opposition in the European Patent Office. Both occur at the post-grant stage.

We show in section 3 that bargaining breakdown is more likely to happen and a challenge ensue when the dispute involves a strong patent, for the patent-holder will be "tougher" at the bargaining table. Private force, then, may be exerted toward the wrong target, and the true inventor may face a higher litigation risk than the opportunistic player.

Even when the weak patent can be eliminated by private challenges, it doesn't necessarily imply that we can rely on private force to such an extent that the patent office should reduce or maintain low examination standards. In section 4, we show that a greater effort at the patent office may *increase* the chance to eliminate the weak patent through court challenges. There may be a positive relationship between public and private enforcement. Together with the case selection pattern, these results cast doubts on the "rational ignorance" hypothesis and call for reforms to improve patent office performance. In a sense, we provide a *raison-d'etre* for the patent office, and refute the idea of abolishing patent office examination and move toward a patent registration system.[2]

In section 5, we introduce two additional policy tools: application fees and a pre-grant challenge procedure. We show that in the two-type case a fee that fully deters the opportunistic player from filing a patent application will crowd out private enforcement, but can't substitute for public enforcement. Concerning a pre-grant challenge system, we point out some of its limitations, including the reversal of case selection pattern and the challenger's choice of timing to initiate a challenge. Section 6 concludes the paper and discusses future research. All proofs are relegated to Appendix A; and Appendix B extends our main results to alternative settings, especially the one where the patent-holder has continuous types.

□ **Related literature:** In law and economics, case selection has been extensively studied under two prominent approaches, that of "divergent expectations" and "asymmetric information".[3] Meurer (1989) provides an application of the asymmetric infor-

[2]See Merges (1999).

[3]A seminal paper using divergent expectations is Priest and Klein (1984). For the asymmetric information paradigm, the theoretical literature has been fairly well developed in several directions. Besides the screening model, where the uninformed party makes the offer (Bebchuk, 1984), there are also studies of: one-sided asymmetric information with the informed party makes the offer (the signaling case); two-sided asymmetric information; and the dynamic multiple-offer bargaining situation, *etc.*. Spier (2005) is a recent review of the literature. On the other hand, most empirical studies use the divergent expectations. But there is no definite evidence supporting either paradigm. Waldfogel (1998) favors the divergent expectations story, while Froeb (1993) supports the asymmetric information approach.

mation paradigm to the patent context.[4] We follow the same approach on the ground that the low patent quality problem can be alleviated through discouraging applications on technologies already in the public domain, a complaint widely shared, among others, in the software industry. A natural modeling strategy is to consider a situation where the patent applicant, but not other parties, is aware of this gaming behavior, and public policy should address this opportunism.

In the patent literature, recent concerns about the patent quality have attracted reform proposals from different sources, such as the United States Federal Trade Commission (FTC 2003), National Academies of Science (2004), as well as numerous law and economics scholars.[5] These reform proposals cover almost all aspects of patent life, from filing of applications, prosecution at the patent office, post-grant challenges, to patent litigation, but often lack sufficient formal analysis. One reason, perhaps, is that relative to the optimal policy design in terms of patent length, scope, and other instruments, very few theoretical efforts have been devoted to patent examination, or more generally the implementation of the patent system. A paper by a law scholar, Kesan (2005), describes how "bad," or weak patents can be settled in a symmetric information environment with legal expenses. On the other hand, two works by economists, Langinier and Marcoul (2003) and Caillaud and Duchêne (2005), elaborate on the patent application strategy and its relationship to patent office examination.

Langinier and Marcoul (2003) considers the patent applicant's search and disclosure of information to the patent office, while the later performs a complementary search and examination upon receiving the applicant's disclosure report. Caillaud and Duchêne (2005) considers multiple firms' R&D and patent filing strategies when the patent office faces the overload problem, that is, when the examination effort upon each application is decreasing due to application volume. For those firms pursuing opportunistic patenting, i.e., seeking patent protection on existing technologies, their applications' survival rate depends on others' strategy, and so multiple equilibria exist: if few file patent applications, then a high level of patent office scrutiny is received by each application; but as more firms "jointly attack" the patent office, an application receives a lower level of examination and a higher survival rate, as a consequence of the resource constraint of the patent office (the overload problem). Different from these papers, we emphasize the "second eye", that is, the role of the private sector in the

[4]But there is no litigation effort choice in his model. As we shall see, this is a crucial element for our results.

[5]Interested readers are referred to the special issue of *Berkeley Technology Law Journal*, 2004, 19 (3).

patent examination process, and consider the interaction between public and private sectors in improving patent quality.[6]

2 Model

There are three players: An inventor A (she) seeks patent protection for her invention, which, if an application is filed, is examined by the patent office (P) and possibly by a private challenger (B, he) in court to verify whether the invention fulfills the patentability requirements specified in the patent law.

Suppose that, under perfect examination, A's application will be rejected with a probability θ. For instance, the patent examination body (say, the patent office) has full access to all relevant information, and with probability θ a piece of patent-defeating prior art exists which proves that A's invention doesn't satisfy one or several of the patentability requirements. This probability is referred to as the "invalidity" of the patent (when issued). For simplicity, consider a two-type case $\theta \in \{\underline{\theta}, \bar{\theta}\}$, with $0 < \underline{\theta} < \bar{\theta} \leq 1$ (the case of $\underline{\theta} = 0$ will be treated in an example). An inventor with low invalidity $\underline{\theta}$, or high validity, is said to be a "true" inventor, or the "good" type: She spends considerable resources in R&D activities and brings about technological breakthrough. By contrast, an inventor with high invalidity $\bar{\theta}$ is called the "bad," or "opportunistic" type: She exploits the public domain and tries to patent an "old" technology. We also refer to a patent with high validity $\underline{\theta}$ as a "strong" patent, and one with $\bar{\theta}$ as a "weak" patent. Assume that θ is the inventor's private information, and other parties hold common initial belief that $Pr(\underline{\theta}) = \alpha$. Define $\theta^0 \equiv \alpha\underline{\theta} + (1-\alpha)\bar{\theta}$ as the *ex ante* average invalidity.

A positive probability to deny the true inventor patent protection, $\underline{\theta} > 0$, may come from a "type II" error in the patent examination process. Patentability standards may be inappropriately interpreted such that, for instance, once an invention is realized, others may perceive it as easier to achieve than it actually was. This "hindersight" bias may render an invention "obvious" or lacking an "inventive step," and so patent protection is denied. Alternatively, the patent authority may grant the monopoly rights to a good inventor only with some probability in order to reduce the deadweight loss

[6]This paper, in a broad sense, is therefore related to another research field in law and economics, namely, the cooperation of private and public sectors in law enforcement. Shavell (1993) discusses the costs and benefits of private enforcement, and the resulting optimal incorporation of private enforcement in different legal fields. This paper illustrates case selection bias as another limitation of private enforcement.

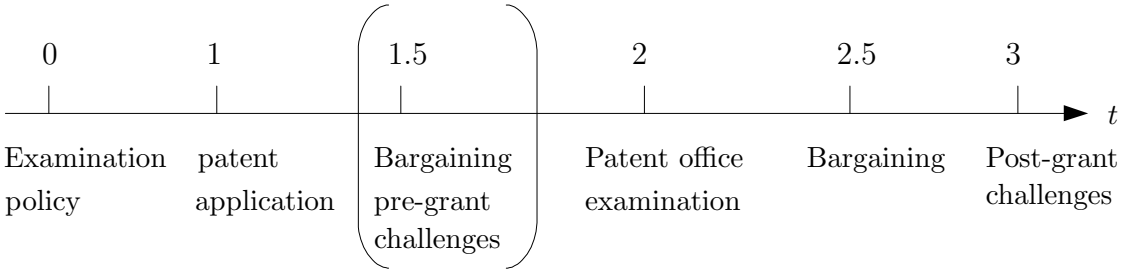


Figure 1: Timing

(Ayres and Klemperer, 1999).

We model patent examination as a "search and destroy" process: P and B can exert costly efforts e_P and e_B, respectively, to search for the prior art, and the patent protection is denied if and only if the defeating prior art is found. Assume that, conditional on the existence of prior art, P's and B's search results are independent of each other. Given $\theta \in \{\underline{\theta}, \bar{\theta}\}$, the probability to eliminate A's application by the patent office (the private challenger) is $\theta \cdot e_P$ ($\theta \cdot e_B$, respectively). The private party B's search cost is $c(e_B)$, with $c(0) = c'(0) = 0$, $c(1) = c'(1) = \infty$, and c' as well as $c'' > 0$. Later we will consider the patent office's cost, and assume that P is less efficient than B. We call e_P (e_B) public (private, respectively) enforcement efforts.

Concerning payoffs, regardless of her type, A gets a monopoly profit $\pi > 0$ when receiving the patent protection, and B gets a benefit $b \in (0, \pi)$ when the patent application is rejected. Otherwise the two receive no return. Except in section 5, private players are protected by limited liability. On the other hand, the patent office is concerned with the patent quality, which, in the two-type case, can be conveniently defined as the probability that the patent is issued to the true inventor. The patent office therefore aims to eliminate as much as possible the likelihood of granting patent rights to the opportunistic inventor, whether through private or public efforts.[7]

We first restrict the patent office's policy tool to examination efforts e_P. We then consider, separately, application fees and the possibility of mounting a private patent challenge at an alternative time, namely the pre-grant stage. We assume that the patent office can commit to its policy. FIGURE 1 illustrates the timing of the game: The patent office first announces its examination policy; and A decides whether to file a patent application based on the policy. Under a post-grant challenge system, a

[7]For most part of the analysis, we ignore the impact of patent examination on the true inventor's returns from using the patent system and so her R&D incentives. See the concluding remark in section 4.

patent application first undergoes the patent office examination, and, upon issuance, encounters a private challenge by B. But the two parties bargain to settle the case before the court fight. On the other hand, under a pre-grant challenge the private enforcement and bargaining take place before the patent office examination. We assume that, when bargaining, A makes a take-it-or-leave-it offer to B. (In APPENDIX B, we show that our main results are robust to the alternative distribution of bargaining power, i.e., when B makes the offer, and a more general setting where A has continuous types.)

3 The Limit of Private Enforcement

In this section we demonstrate that under a post-grant challenge system, a case involving a weak patent ($\bar{\theta}$) is more likely to be settled than that involving a strong patent ($\underline{\theta}$). This pattern of case selection points out the limit of private enforcement, and is key to subsequent analysis.

Suppose that B's litigation effort e_B is not contractible and so cannot be part of the settlement agreement.[8] A settlement offer is a transfer between A and B. Let $\hat{\alpha} \in (0,1)$ be the belief that B faces a good inventor at the beginning of the bargaining subgame. This probability is affected by the patent office examination effort e_P and can be seen as the quality of an issued patent. Define $\hat{\theta} \equiv \hat{\alpha}\underline{\theta} + (1-\hat{\alpha})\bar{\theta}$ and the following terms: with $\theta \in \{\underline{\theta}, \bar{\theta}\}$,

$$e_B^*(\hat{\theta}) \equiv \arg\max_{e_B} \ \hat{\theta} e_B b - c(e_B),$$
$$u_A(\theta, e_B^*) = (1 - \theta e_B^*)\pi, \quad \text{and} \quad u_B(\hat{\theta}) = \hat{\theta} e_B^* b - c(e_B^*).$$

e_B^* is B's optimal litigation effort, and u_A and u_B are A's and B's expected payoffs in litigation, respectively. The optimal litigation effort is increasing in $\hat{\theta}$, and so decreasing in $\hat{\alpha}$. A lower probability to find the information and strike down the patent discourages B's search activity. On the other hand, when engaging in a legal fight, A always prefers a less intensive attack from B, i.e., a lower e_B^*, while B's payoff is increasing in the probability of facing a weak patent $\hat{\theta}$.

Denote $\underline{e}_B \equiv e_B(\underline{\theta})$ and $\bar{e}_B \equiv e_B(\bar{\theta})$, and so $e_B^* \in [\underline{e}_B, \bar{e}_B]$. Note that $\underline{e}_B > 0$ for $\underline{\theta} > 0$. It is easy to check that $u_A(\underline{\theta}, e_B) > u_A(\bar{\theta}, e_B)$, $\forall e_B \in [\underline{e}_B, \bar{e}_B]$, and $u_A(\theta, e_B^*)$

[8] This effort may not be observable. Even if observable, the court may not enforce an agreed effort level to be exerted in litigation.

is increasing in $\hat{\alpha}$. That is, given the same private litigation effort, the true inventor's expected payoff from litigation is strictly higher than that of the opportunistic player; and through its effect on e_B^* via $\hat{\theta}$, an inventor's litigation payoff is increasing in the belief $\hat{\alpha}$. Also note that by $b < \pi$, the case is always settled under symmetric information: $\pi - u_B(\underline{\theta}) > u_A(\underline{\theta}, \underline{e}_B)$ and $\pi - u_B(\bar{\theta}) > u_A(\bar{\theta}, \bar{e}_B)$.

PROPOSITION 1. *(Case selection) After patent issuance, whether A or B makes a take-it-or-leave-it offer, there is no bargaining equilibrium in which only the true inventor settles.*

This result is fairly general and well-established in the literature of law and economics, regardless of the distribution of bargaining power. Intuitively, when one party holds private information about her case quality (θ here), a stronger case (lower θ) makes a "tougher" player on the bargaining table, and so a settlement deal is harder to reach.

We now consider when private enforcement can be mounted against a weak patent. The weak patent is said to be fully (partially) exposed to private enforcement if the opportunistic A engages in litigation for sure (with a probability, respectively). By PROPOSITION 1, whenever the opportunistic A litigates, so does the good A.

PROPOSITION 2. *(Private enforcement) Suppose that A makes the settlement offer. The weak patent is subject to private enforcement when $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$. Suppose this is true.*

- *(Full exposure) When $u_A(\bar{\theta}, e_B^*(\hat{\theta})) \geq \pi - u_B(\bar{\theta})$, there is a Perfect Bayesian Equilibrium (henceforth, PBE) in which no settlement is reached at all, and B exerts litigation effort $e_B^*(\hat{\theta})$; and*
- *(partial exposure) if $u_A(\bar{\theta}, e_B^*(\hat{\theta})) < \pi - u_B(\bar{\theta}) < u_A(\bar{\theta}, \underline{e}_B)$, there is a PBE in which the opportunistic A litigates with probability $x^* \in (0,1)$, the good A always litigates, and B, with a belief α_x^* upon litigation, exerts an litigation effort $e_{B,x}^* < e_B^*(\hat{\theta})$, where $e_{B,x}^*$, x^*, and α_x^* are determined by*

$$u_A(\bar{\theta}, e_{B,x}^*) = \pi - u_B(\bar{\theta}),\ e_{B,x}^* = e_B^*\big(\alpha_x^*\underline{\theta} + (1-\alpha_x^*)\bar{\theta}\big),\ \text{and}\ \alpha_x^* = \frac{\hat{\alpha}}{\hat{\alpha} + (1-\hat{\alpha})x^*}. \quad (1)$$

By this proposition, private enforcement can possibly eliminate the weak patent only when $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$. To understand this condition, note that $u_A(\bar{\theta}, \underline{e}_B)$ and $u_B(\bar{\theta})$ are the opportunistic A's and B's highest possible payoff in litigation, respectively, and so offering these amounts to corresponding players guarantees acceptance.

Suppose that $u_A(\bar{\theta}, \underline{e}_B) \leq \pi - u_B(\bar{\theta})$. When A makes the offer, the opportunistic A's highest possible litigation payoff is smaller than the lowest possible payoff from settlement, which is obtained by offering B's highest litigation payoff to ensure settlement. She then has every incentive to settle.[9] In this case, private force is either exerted toward the wrong target (the strong patent), or simply not active; and patent quality cannot be improved by private enforcement.

COROLLARY 1. *When $u_A(\bar{\theta}, \underline{e}_B) \leq \pi - u_B(\bar{\theta})$, private enforcement doesn't improve the patent quality. It reduces the quality of issued patents when only the good patent-holder engages in litigation.*

REMARK 1. (EQUILIBRIUM REFINEMENT) In the proof we show that these equilibria survive the criterion $D1$ (Cho and Kreps, 1987). This criterion constrains the weight B can put on the opportunistic type upon the off-path event of litigation. Roughly speaking, the good A would have more to gain than the opportunistic A in a legal fight, and so $D1$ requires the opportunistic A be fully deleted from B's off-path belief.

In the proof of PROPOSITION 2, we also consider other bargaining outcomes such as where both types of A settle and there is no litigation, and where only the good A litigates. However, no PBE exists that fulfills the criterion $D1$ and implements the two outcomes.[10] ■

REMARK 2. ("HARASSING" THE TRUE INVENTOR) The case selection pattern also implies a higher litigation risk for the true inventor, which may translate into a higher probability to lose the patent protection. This happens when B litigates only against the good A, while settling the case with the opportunistic A. In other words, a true inventor may be "harassed" when trying to enforce her patent rights.[11] Private enforcement, then, may reduce the true inventor's payoff and impair R&D incentives without offsetting gains to raise the patent quality. ■

Before proceeding to the relationship between public and private enforcement, let us consider two special cases of private bargaining.

---

[9] In APPENDIX B we show that the same condition applies when B makes the offer.

[10] "Divinity," though, retains these bargaining outcomes (Bank and Sobel, 1987). It is a weaker than $D1$ and only requires that B believe the good A plays the deviant move at least as often as the opportunistic A. The "passive belief," for example, is allowed under divinity but not under $D1$.

[11] The harassment hypothesis usually refers to invalidation challenges facing a patent-holder from potential infringers or other stake-holders. One possible litigation shown in our model is exactly this invalidation suit.

EXAMPLE 1. (AN IRONCLAD GOOD PATENT) When the good patent can never be invalidated, $\underline{\theta} = 0$, the opportunistic A can still be subject to private litigation. This is confirmed by that fact that, under this case, $u_A(\bar{\theta}, \underline{e}_B) = \pi > \pi - u_B(\bar{\theta})$.

However, without invalidation risk the true inventor will never pay B to settle the case, there is no equilibrium in which private bargaining always reaches a deal, whoever makes the offer. Another equilibrium outcome ruled out by this assumption is one in which B learns A's true type and settles with the opportunistic player while litigating against the true inventor. By $\underline{\theta} = 0$ and so $\underline{e}_B = 0$, this equilibrium is busted by the opportunistic A's attempt to mimic the good type (and engage in a "legal fight" with no litigation efforts from B). ■

EXAMPLE 2. (INELASTIC PRIVATE ENFORCEMENT CAPACITY) Suppose that $\underline{\theta} > 0$ but B has inelastic litigation capacity. For simplicity, let us consider the extreme case of fixed and costless $e_B > 0$.[12] After this modification, the weak patent is entirely exempted from private enforcement. A fixed e_B renders $u_B(\bar{\theta}) = \bar{\theta} e_B b < \pi - u_A(\bar{\theta}, e_B) = \bar{\theta} e_B \pi$, which violates the necessary condition $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$.[13] This confirms that B's litigation effort decision is a key ingredient in our analysis. ■

4 Public vs. Private Enforcement

The results we obtain in the previous analysis cast doubts on Lemley (2001)'s hypothesis of a "rationally ignorant patent office." Since private force cannot only be directed toward the "right target," that is, the weak patent, provoking private litigation at best improves the patent quality at the expense of the true inventor, who suffers from burdensome litigation and lower return from R&D.

Even if concerns about innovation incentives are not present, a closer look at PROPOSITION 2 shows that a proposal to reduce or to maintain a low level of patent office examination may be detrimental to the overall patent quality control standard. This section develops the relationship between public and private enforcement, through the former's effect on the patent quality $\hat{\alpha}$. For simplicity, suppose that the good inventor's R&D incentives are not too much damaged during the patent-granting process. For instance, we may assume that $\underline{\theta} > 0$ but low enough so that even if $e_P = e_B = 1$,

[12] With costly but fixed effort, we need only that B has a credible threat to incur the cost in a legal fight, e.g., by assuming a cost smaller than $\underline{\theta} e_B b$.

[13] Introducing litigation cost only strengthens this result.

the expected return from patenting, $(1-\underline{\theta})^2\pi$, covers her R&D expenditure.

Recall that $\theta^0 \equiv \alpha\underline{\theta}+(1-\alpha)\bar{\theta}$. When the patent office exerts an examination effort $e_P \geq 0$, the quality of an issued patent is

$$\hat{\alpha} = \frac{\alpha(1-\underline{\theta}e_P)}{\alpha(1-\underline{\theta}e_P)+(1-\alpha)(1-\bar{\theta}e_P)} = \frac{\alpha(1-\underline{\theta}e_P)}{1-\theta^0 e_P} \tag{2}$$

$$\Rightarrow \quad \frac{\partial\hat{\alpha}}{\partial e_P} = \frac{\alpha(1-\alpha)(\bar{\theta}-\underline{\theta})}{\{1-\theta^0 e_P\}^2} > 0.$$

A higher level of public enforcement raises the patent quality. Next, suppose that $\pi - u_B(\bar{\theta}) < u_A(\bar{\theta}, \underline{e}_B)$ and so the weak patent can be subject to private enforcement. We consider the full and partial exposure regime in turn, i.e., whether the opportunistic patent-holder litigates with probability equal to or less than one.

The full exposure regime requires patent quality $\hat{\alpha}$ be high enough, so that $\hat{\theta}$ and litigation effort e_B^* low enough: $u_A(\bar{\theta}, e_B^*(\hat{\theta})) \geq \pi - u_B(\bar{\theta})$.[14] Intuitively, the opportunistic inventor is willing to mix with the good inventor and litigate only when she expects to encounter a low litigation effort. This is more likely to be the case when patent office exerts great examination effort e_P and maintains high patent quality $\hat{\alpha}$. In addition, in this regime a marginal increase in public enforcement e_P will *reduce* private enforcement effort e_B, for a higher patent quality $\hat{\alpha}$ weakens B's search intensity. In other words, in this regime public enforcement crowds out private enforcement.

The partial exposure regime, on the other hand, happens for low $\hat{\alpha}$.[15] This regime exhibits an interesting relationship between public and private enforcement. By Proposition 2 the opportunistic A's litigation probability $x^* = [\hat{\alpha}(1-\alpha_x^*)]/[(1-\hat{\alpha})\alpha_x^*]$ is *increasing* in $\hat{\alpha}$. Together with the fact that the belief α_x^* and litigation effort $e_{B,x}^*$ are fixed in this case, the probability that the weak patent will be eliminated by private force, $x^* \cdot e_{B,x}^*$, is also *increasing* in e_P. Different from the full exposure regime, here public enforcement *crowds in* private enforcement.[16]

The reason is, referring to condition (1), under partial exposure the litigation effort $e_{B,x}^*$ is determined such that the opportunistic A is indifferent between paying $u_B(\bar{\theta})$ to settle the case and facing a challenge with effort $e_{B,x}^*$. On the other hand, to have $e_{B,x}^*$ as the best response, B should have a belief α_x^* when filing a challenge. And since

[14] If B makes the offer, by contrast, full expose happens only when $\hat{\alpha}$ is small enough (and A accepts the offer upon indifference). Nevertheless, this is only the qualitative difference between the two distributions of bargaining power. See Appendix B.

[15] This is also true when B makes the offer.

[16] The same holds true when B makes the offer, provided that $\hat{\alpha}$ is low enough and B's cost function is well-behaved.

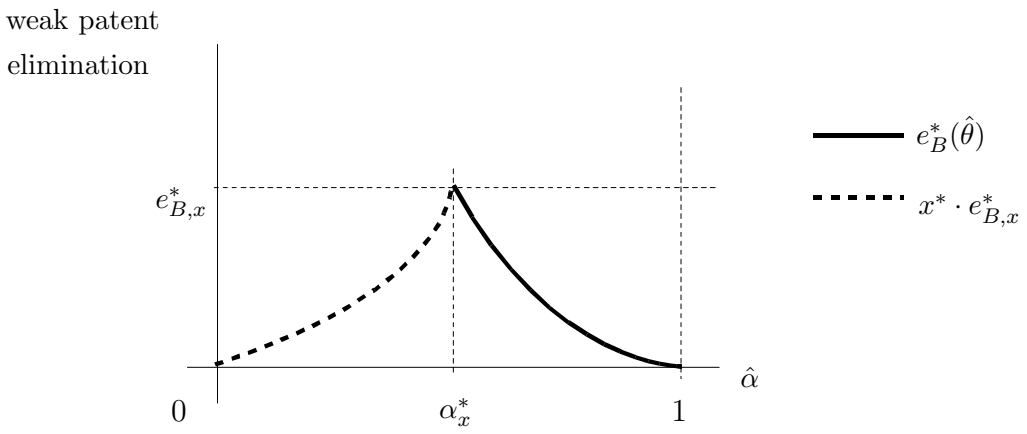


Figure 2: Patent quality and private enforcement

a higher e_P will raise the quality of an issued patent $\hat{\alpha}$, the opportunistic A should litigate more (raise x^*) in order to fix B's belief at α^*_x.

PROPOSITION 3. *(Public and private enforcement) Assume $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$ so that the weak patent may be subject to private enforcement.*

- *(Full exposure) When $\hat{\alpha} \geq \alpha^*_x$, the weak patent is litigated for sure, and an higher level of public enforcement e_P crowds out private litigation efforts $e^*_B(\hat{\theta})$.*
- *(Partial exposure) When $\hat{\alpha} < \alpha^*_x$, the weak patent is litigated with probability x^*, and the probability to eliminate a weak patent through private effort, $x^* \cdot e^*_{B,x}$ is increasing in e_P, even though B's litigation efforts $e^*_{B,x}$ is not affected by e_P.*

FIGURE 2 summarizes the impact of patent quality $\hat{\alpha}$ on "weak patent elimination," which is defined as the probability that the weak patent will be eliminated in litigation. (Since $\hat{\alpha}$ is strictly increasing in e_P, it also depicts the effect of public enforcement on private enforcement.) When the patent quality increases, we move from the partial exposure (the dashed line) to the full exposure regime (the solid line). A marginal increase in the patent quality raises the probability of eliminating the weak patent in the former case, but not in the latter case. There is a non-monotonic relationship between weak patent elimination and the patent quality.

Notice the policy implication. A positive relationship between public enforcement and weak patent elimination occurs precisely under low patent quality. As previously discussed in the introduction, the current debate about patent quality is centered on the complaint that the patent office has issued too many unwarranted patents. To

address this concern, we may want to improve the performance of the patent office not only to directly raise the patent quality, but also to enhance the involvement of private force in the quality control process.

□ **When to reduce public enforcement?** One might wonder, given a negative relationship between public and private enforcement at the full exposure regime, when the private challenger enjoys a cost advantage over the public agency, we should constrain the patent office examination and let a patent be scrutinized later through private litigation efforts. In other words, Lemley's "rational ignorance" hypothesis might be vindicated in this case.

To check this possibility, we assume that the patent office has a cost function $\gamma c(e_P)$, where $\gamma \geq 1$ and $c(\cdot)$ is B's cost function. Define the total cost of patent examination as $C(e_P) \equiv \gamma c(e_P) + (1-\theta^0 e_P)c(e_B^*(\hat{\theta}))$. Also define the level of examination a patent application is expected to receive as $e_P + e_B$, for under this regime, a patent applicant with θ expects rejection with probability $1-(1-\theta e_P)(1-\theta e_B) = \theta(e_P+e_B)-\theta^2 e_P e_B \simeq \theta(e_P+e_B)$. We show when a marginal reduction in e_P will reduce the total cost without deteriorating the examination standard.

A marginal change in e_P causes a change in examination standards by

$$\frac{d[e_P + e_B^*(\hat{\theta})]}{de_P} = 1 + \frac{de_B^*(\hat{\theta})}{d\hat{\alpha}}\frac{\partial \hat{\alpha}}{\partial e_P} = 1 - \frac{(\bar{\theta}-\underline{\theta})b}{c''(e_B)}\frac{\alpha(1-\alpha)(\bar{\theta}-\underline{\theta})}{(1-\theta^0 e_P)^2},$$

and a change in the total cost by

$$\frac{dC(e_P)}{de_P} = \gamma c'(e_P) - \theta^0 c(e_B^*(\hat{\theta})) + (1-\theta^0 e_P)c'(e_B^*(\hat{\theta}))\frac{de_B^*(\hat{\theta})}{d\hat{\alpha}}\frac{\partial \hat{\alpha}}{\partial e_P}.$$

The following result is obtained from these two expressions in a straightforward manner.

Proposition 4. *(A rationally ignorant patent office under the full exposure regime) Under the full exposure regime, a marginal decrease in e_P does not weaken the overall examination standard if and only if*

$$\frac{de_B^*(\hat{\theta})}{d\hat{\alpha}}\frac{\partial \hat{\alpha}}{\partial e_P} \leq -1 \quad \Rightarrow \quad \frac{\alpha(1-\alpha)(\bar{\theta}-\underline{\theta})^2 b}{c''(e_B)(1-\theta^0 e_P)^2} \geq 1, \tag{3}$$

and reduces the total examination cost if and only if

$$\gamma > \frac{1}{c'(e_P)}\left[\theta^0 c(e_B^*(\hat{\theta})) - (1-\theta^0 e_P)c'(e_B^*(\hat{\theta}))\frac{de_B^*(\hat{\theta})}{d\hat{\alpha}}\frac{\partial \hat{\alpha}}{\partial e_P}\right]. \tag{4}$$

The rational ignorance hypothesis is supported when both conditions hold.

Not surprisingly, the private sector's cost advantage γ should be large enough to justify a not-so-excellent patent office. In the proof of this proposition we also obtain a sufficient condition for condition (3) to hold: $\forall e_B$, $\alpha(1-\alpha)(\bar{\theta}-\underline{\theta})^2 b \geq c''(e_B)$. This stems from the fact that the private sector's response should be large enough in order to compensate for a more lax public quality control. Among others, this requires a "less curved" cost function, i.e., c'' small enough, as $\partial e_B^*(\hat{\theta})/\partial\hat{\theta} = b/c''(e_B^*)$.

REMARK. (R&D INCENTIVES) So far we've ignored the true inventor's R&D incentives. If this concern is introduced, to restrain the magnitude of type II error the patent office may want to constrain its examination effort e_P. However, previous analysis shows that a marginal reduction in e_P may not always decrease the overall examination effort. This is indeed true in the partial exposure regime. In the full exposure regime, a reduction in e_P causes e_B to increase, and the general enforcement level decreases if and only if condition (3) fails. ■

5 Other Policy Choices

In this section we first erase the limited liability protection and allow negative returns for an inventor. This allows us to introduce applications fees as an additional policy tool. We then turn to an alternative timing to exert private efforts, i.e., a pre-grant challenge system.

□ **Application fees:** When the patent office can charge application fees, this may deter, ideally, the opportunistic inventor from seeking patent protection. In general, to achieve this goal, a more effective way is to condition the pecuniary punishment on the examination outcome, e.g., upon the rejection of a patent application or invalidation of an issued patent in court. However, a fine after invalidation is arguably under the discretion of the court, and an applicant, especially a "short-run player," might simply run away when her application is rejected by the patent office. Instead, we consider a uniform application fee f for all patent applications. Nevertheless, our main result is robust to the exact shape of the pecuniary mechanism.

Suppose that an application fee f fully deters the opportunistic inventor from applying for a patent, but not the true inventor. When this is true, at the bargaining stage B holds belief that $\hat{\alpha} = 1$, and symmetric information prevents bargaining breakdown. In this two-type case, a fully deterrent application fee mutes entirely private

enforcement. When A holds the bargaining power,[17] it suffices to pay $u_B(\underline{\theta})$ to settle the case, and a deterrent fee f should satisfy

$$(1-\bar{\theta}e_P)\pi - u_B(\underline{\theta}) < f \leq (1-\underline{\theta}e_P)\pi - u_B(\underline{\theta}).$$

Since this condition will not hold $e_P = 0$, a deterrent application fee cannot substitute for patent office examination. Furthermore, to preserve the good inventor's R&D incentives, the patent office should set f as small as possible, without losing its deterrent power. Let $f^D = (1-\bar{\theta}e_P)\pi - u_B(\underline{\theta}) + \epsilon$, with $\epsilon > 0$ but small. Since f^D is decreasing in e_P, the good inventor's payoff, $(1-\underline{\theta}e_P)\pi - u_B(\underline{\theta}) - f^D = (\bar{\theta}-\underline{\theta})e_P\pi - \epsilon$, s increasing in e_P:

PROPOSITION 5. *(Application fees) In the two-type case, an application fee that fully deters opportunistic patenting crowds out private enforcement but cannot substitute for public enforcement. A higher patent office examination level e_P reduces the necessary fee. And when the application fee is set at the minimal necessary level f^D, the good inventor's payoff, and so the R&D incentive, is increasing in e_P.*

□ **Pre-grant challenges:** Lastly, let us consider a pre-grant challenge system. Suppose that after receiving a patent application but before starting its examination process (time 1.5 in FIGURE 1), the patent office publishes the application and allows third parties to challenge it (or submits information concerning its patentability).[18]

This alternative timing of the challenge allows the patent office to set different examination levels according to an application's history. Let e_P^c be the examination effort exerted on an application that has survived private challenges, and e_P^n on that which has not yet been challenged. Intuitively, the patent office should set $e_P^c \leq e_P^n$. In addition to the reason that private enforcement efforts perform as a "certificate" about the validity of an application, case selection (PROPOSITION 1) provides further support of such a policy, because a weak patent (application) is less likely to receive private scrutiny.

[17] The distribution of bargaining power is not crucial to this result. It only changes the level of f to deter opportunistic patenting, for the patent-holder's payoffs from fully settling the case depend on who makes the offer.

[18] Early publication of patent applications (18 months after filing) has been widely adopted in Japan and Europe; the U.S. has the same procedure but allows an applicant to opt out. About the pre-grant challenge, the 2007 Patent Reform Act in the U.S. introduces a procedure permitting third parties to submit relevant information before the issuance of a patent.

However, under such a policy, an applicant may try to circumvent the high effort e_P^n by arranging a "fake" challenge, in particular when the patent office is unable to verify the challenger's effort level, that is, whether the challenger only initiates a nominal challenge procedure without any serious effort to strike down the application. Besides, we show that (i) the "direction" of case selection may be reversed at the pre-grant challenge stage. That is, contrary to the previous result, there may exist an equilibrium where only the true inventor settles at the pre-challenge bargaining; and (ii) when B does intend to initiate a challenge, and both pre- and post-grant challenges are available, he may want to wait and file a private challenge only after the failure of the patent office.[19]

For the first point, suppose that B can only initiate a challenge at the pre-grant stage, and that A's settlement payment comes from the monopoly rent and so is paid only when the patent is issued. (This is the case when A is protected by limited liability.) Recall that B cannot commit to e_B in an agreement, and his initial belief of patent (application) quality is α. We derive conditions under which there is a separating equilibrium where only the good inventor settles. A necessary condition is both $\underline{\theta}$ and $e_P^c > 0$. The former is simply due to the fact that a true inventor with $\theta = 0$ will never pay anything to settle. The latter can be justified in that the patent office doesn't "outsource" the examination task entirely to private parties.[20] Even if an application survives private challenges, the patent office still does its own work.

Intuitively, when the patent office sets different examination levels according to the challenge history, A will take this into account when making settlement decisions. Consider if $e_P^n >> e_P^c$, that is, if an unchallenged application will receive a more detailed examination than an application surviving private challenges. This gives an applicant incentives not to settle with a private challenger in order to avoid stringent public scrutiny. But the magnitude of this effect depends on the true quality of the invention θ. For instance, when $\underline{\theta}$ is very close to zero, even $e_P^n \simeq 1$ won't harm the true inventor too much. The case selection pattern at the pre-grant challenge stage may be reversed. That is, only the good A settles while the opportunistic A experiences a private challenge. The following proposition confirms this possibility.

PROPOSITION 6. *(Pre-grant challenges and reverse case selection) Suppose that B can*

[19] Of course, this is more likely the case when costs accrued to challengers are not so different for the post- and pre-grant challenge procedures.

[20] Or, equivalently, the patent office doesn't "rubber stamp" the issuance of a patent following private efforts.

only file a challenge at the pre-grant stage. There is a PBE where only the opportunistic A is challenged when

$$\frac{(1-\bar{\theta}\bar{e}_B)(1-\bar{\theta}e_P^c)}{(1-\bar{\theta}e_P^n)} \geq \frac{\pi-\underline{s}}{\pi} \geq \frac{(1-\underline{\theta}\bar{e}_B)(1-\underline{\theta}e_P^c)}{(1-\underline{\theta}e_P^n)}, \tag{5}$$

where $\underline{s} = [u_B(\underline{\theta}) + (1-\underline{\theta e}_B)\underline{\theta}e_P^c b]/(1-\underline{\theta}e_P^n)$.

First note that condition (5) won't hold when $e_P^c = 0$. In this case, a necessary condition of this equilibrium,

$$\frac{(1-\bar{\theta}\bar{e}_B)(1-\bar{\theta}e_P^c)}{(1-\bar{\theta}e_P^n)} \geq \frac{(1-\underline{\theta}\bar{e}_B)(1-\underline{\theta}e_P^c)}{(1-\underline{\theta}e_P^n)},$$

reduces to $e_P^n \geq \bar{e}_B$, contradictory with

$$\frac{1-\underline{\theta}\bar{e}_B}{1-\underline{\theta}e_P^n} \leq \frac{\pi-\underline{s}}{\pi} < 1.$$

In order to consider when it's more likely to have this equilibrium, let us fix $\bar{e}_B$, $\underline{\theta}$, and e_P^c at strictly positive levels, but less than one. Suppose that $\underline{s}$ is small enough (due to, say, a small b) so that

$$\frac{\pi-\underline{s}}{\pi} \geq (1-\underline{\theta}\bar{e}_B)\frac{1-\underline{\theta}e_P^c}{1-\underline{\theta}} \geq (1-\underline{\theta}\bar{e}_B)\frac{1-\underline{\theta}e_P^c}{1-\underline{\theta}e_P^n}.$$

That is, the second inequality in condition (5) holds for all e_P^n. In this case, the separating equilibrium exists as long as

$$(1-\bar{\theta}\bar{e}_B)\frac{1-\bar{\theta}e_P^c}{1-\bar{\theta}e_P^n} \geq 1 \Rightarrow \frac{1-\bar{\theta}e_P^c}{1-\bar{\theta}e_P^n} \geq \frac{1}{1-\bar{\theta}\bar{e}_B}.$$

For all possible $\bar{\theta}$, it is more likely to hold as e_P^n grows larger. In the extreme case of $\bar{\theta} = 1$, this condition is guaranteed when e_P^n is large enough. This equilibrium exists exactly when the weak patent is of the worst kind, and the patent office exerts maximal efforts to eliminate it with the information provided by case selection!

Remark. (Can sequential private challenges reverse the pattern?) One might suspect that this reverse pattern of case selection is generated by sequential efforts to eliminate patent applications, and could happen as well under post-grant challenges and multiple potential challengers.

For simplicity, suppose there are two potential challengers B_1 and B_2, with identical cost $c(\cdot)$ and benefit b . If A's bargaining with B_1 results in the litigation of opportunistic A and settlement of good A, then B_1 exerts litigation efforts $\bar{e}_B$. Denote the good

A's settlement offer as s. This separating equilibrium fully reveals A's type, and so, knowing the litigation history, there will be no litigation between B_2 and A (when the opportunistic A survives B_1's challenge). B_2 will settle with the good (opportunistic) A with a payment $u_B(\underline{\theta})$ ($u_B(\bar{\theta})$, respectively). Since

$$\pi - s - u_B(\underline{\theta}) \geq (1 - \underline{\theta}\bar{e}_B)\pi - u_B(\bar{\theta}) > (1 - \bar{\theta}\bar{e}_B)\pi - u_B(\bar{\theta}),$$

the opportunistic A will deviate to mimic the good A. The reverse pattern of case selection will not happen under sequential private challenges. ■

Now, consider a potential challenger's timing choice. Suppose that both pre- and post-grant challenges are available to B, but there is only one challenge opportunity. In the absence of a settlement agreement, with belief α and corresponding θ^0,[21] B's payoff from initiating a pre-grant challenge is $u_B(\theta^0)+[1-\theta^0 e_B^*(\theta^0)]e_P^c\theta^0 b$. If B waits after the patent issuance, his expected payoff is $\theta^0 e_P^n b+(1-\theta^0 e_P^n)u_B(\hat{\theta})$, where $\hat{\theta} = \hat{\alpha}\underline{\theta}+(1-\hat{\alpha})\bar{\theta}$ and $\hat{\alpha}$ is determined according to condition (2), with $e_P = e_P^n$. Since $\hat{\alpha} > \alpha$ for all $e_P^n > 0$, $\hat{\theta} < \theta^0$, $e_B^*(\theta^0) > e_B^*(\hat{\theta})$, and $c(e_B^*(\theta^0)) > c(e_B^*(\hat{\theta}))$. We should expect more intensive private challenge efforts at the pre-grant stage than at the post-grant stage.

Since

$$\begin{aligned} u_B(\theta^0) + [1 - \theta^0 e_B^*(\theta^0)]e_P^c\theta^0 b &< \theta^0 e_B^*(\theta^0)b - (1 - \theta^0 e_P^n)c(e_B^*(\theta^0)) + [1 - \theta^0 e_B^*(\theta^0)]e_P^c\theta^0 b \\ &= -(1 - \theta^0 e_P^n)c(e_B^*(\theta^0)) + b\Big[\theta^0 e_B^*(\theta^0) + (1 - \theta^0 e_B^*(\theta^0))\theta^0 e_P^c\Big], \end{aligned}$$

and

$$\theta^0 e_P^n b + (1 - \theta^0 e_P^n)u_B(\hat{\theta}) = -(1 - \theta^0 e_P^n)c(e_B^*(\hat{\theta})) + b\Big[\theta^0 e_P^n + (1 - \theta^0 e_P^n)\hat{\theta} e_B^*(\hat{\theta})\Big],$$

a sufficient condition for B to choose the post-grant procedure is

$$e_P^n - e_P^c > e_B^*(\theta^0)(1 - \theta^0 e_P^c). \tag{6}$$

It is more likely as e_P^n gets larger and e_P^c gets smaller. That is, B will postpone and free ride on public efforts if the patent office targets and exert much higher efforts towards those applications not being protested by private players.

PROPOSITION 7. *(Choice of challenge timing) When condition (6) holds, a potential challenger prefers to challenge at the post-grant stage.*

[21] This α may be the initial belief when there is no bargaining at all between A and B, or the belief after the breakdown of a settlement negotiation.

6 Concluding Remarks

The limitation of private enforcement emphasized in this paper, namely the settlement bias toward weak patents, would persist despite the private challenger's information and cost advantages. These results highlight the importance of a patent office. Accordingly, future theoretical works and reform efforts should figure out how to improve the performance of the patent office in order to "get things right" in the first place. The agency problem and task allocations within the patent office are additional topics in our research agenda.[22]

In this aspect, our analysis sheds some lights on the design of incentive payments for patent examiners. One difficulty in the construction of such an incentive scheme is to find a proper index of examiners' efforts. A straightforward and somewhat "naive" application of incentive theory might suggest the use of court rulings as a measure of performance. A patent examiner would be punished if a patent issued by her is later invalidated in court. Several practical issues reduce the usefulness of this measure: the rare occurrence of patent disputes and the strong tendency toward settlement; upon dispute, the long delay from patent issuance to the final court judgment; and, at least in the United States, a significant portion of patent examiners who choose a career path in the private sector after a few years' experience in the patent office. Our analysis points out another restriction: the information content of a court ruling may be distorted by private bargaining. For instance, a positive relationship between public and private enforcement in the partial exposure regime suggests that a higher effort by the patent examiner may result in more patents being litigated and invalidated in court. It would then be undesirable to punish the examiner upon a successful post-grant court challenge.

Appendix

(To be revised)

A Proofs

□ Proposition 1

[22]Merges (1999) argues that the U.S. patent examiners are given incentives to approve, but not reject patent applications.

Proof. Consider an equilibrium in which the good inventor settles (with some probability) but the opportunistic inventor always litigates. Let s' be (one of) the good inventor's equilibrium settlement payment(s), which may be adopted for some probability, and $e'_B > 0$ be (one of) the litigation efforts facing the opportunistic inventor. When the good inventor prefers settlement and paying s' than litigation against an effort e'_B, $\pi - s' \geq u_A(\underline{\theta}, e'_B) > u_A(\bar{\theta}, e'_B)$, the opportunistic inventor has incentives to deviate to s' and settle. *Q.E.D.*

LEMMA 1. *(Off-path belief selection and full settlement) Consider a PBE where no litigation occurs, and denote s as the equilibrium settlement payment from A to B. If this equilibrium fulfills the criterion $D1$ (divinity), it must be supported by off-path beliefs $\tilde{\alpha} = Pr(\underline{\theta}|\tilde{s})$ such that for $\tilde{s} < s$, $\tilde{\alpha} = 1$ ($\tilde{\alpha} \geq \hat{\alpha}$, respectively).*

Proof. To use $D1$ or divinity to eliminate or constrain the weight on the opportunistic type upon observing a deviation $\tilde{s} < s$, we need to show that whenever a (mixed strategy) best response of B to that deviation makes the opportunistic A (weakly) better off than under the equilibrium, the same best response must give the good A a strictly higher payoff than the equilibrium payoff.

Let s be the equilibrium payment from A to B for a PBE where no litigation ensues. Note that there can be only one such payment, otherwise the player making the offer will deviate to the payment that serves best his/her interests without intriguing law suits. A's equilibrium payoff is $\pi - s$, regardless of her type. Consider B's belief upon an off-path offer $\tilde{s} < s$.

When A makes the offer, upon observing $\tilde{s} < s$, denote B's mixed strategy best response as $(\tilde{\phi}, \tilde{e}_B)$ and belief as $\tilde{\alpha}$, where $\tilde{\phi}$ is the probability to accept the offer and $\tilde{e}_B = e^*_B(\tilde{\theta})$ the litigation effort when rejecting the offer, given $\tilde{\theta} = \tilde{\alpha}\underline{\theta} + (1 - \tilde{\alpha})\bar{\theta}$. A's payoff from deviating to $\tilde{s}$ is therefore $\tilde{\phi}(\pi - \tilde{s}) + (1 - \tilde{\phi})u_A(\theta, \tilde{e}_B)$, $\theta \in \{\underline{\theta}, \bar{\theta}\}$. Note that by the shape of the cost function $c(\cdot)$, B's best response doesn't mix among different values of e_B.

Since $\pi - \tilde{s} > \pi - s$, when $\tilde{\phi} = 1$ both types of A strictly prefer to deviate to $\tilde{s}$. When $\tilde{\phi} = 0$, for any $\tilde{e}_B > 0$, $u_A(\underline{\theta}, \tilde{e}_B) > u_A(\bar{\theta}, \tilde{e}_B)$ and so whenever the opportunistic inventor is (weakly) better off by deviating to $\tilde{s}$, the good inventor strictly prefers doing so. The same holds when $\tilde{\phi} \in (0, 1)$.

When B makes the offer, to support this equilibrium A must reject $\tilde{s}$ and this deviant offer must lead to litigation. Previous argument guarantees that if the opportunistic inventor weakly prefers to deviate under some $\tilde{e}_B$, the good inventor must

strictly prefer doing so. *Q.E.D.*

□ Proposition 2

Proof. Similar to the reason there is at most one equilibrium offer leading to settlement for sure, there can be at most one equilibrium litigation effort e_B.

◇ Full exposure: Along the equilibrium path, both types of A propose a settlement offer $s < u_B(\hat{\theta})$ and B rejects this offer while maintaining belief at $\hat{\theta}$, and so the litigation effort is $e_B^*(\hat{\theta})$. A's equilibrium payoff is $u_A(\theta, e_B^*(\hat{\theta}))$, $\theta \in \{\underline{\theta}, \bar{\theta}\}$. To prevent deviation, (i) since B will agree to settle with a payment $u_B(\bar{\theta})$, the opportunistic A should prefer litigation to settlement for sure, $u_A(\bar{\theta}, e_B^*(\hat{\theta})) \geq \pi - u_B(\bar{\theta})$; and (ii) for other deviations $\tilde{s} < u_B(\bar{\theta})$, B needs to reject $\tilde{s}$ and litigates with $\tilde{e}_B \geq e_B^*(\hat{\theta})$, to be supported by off-path belief $\tilde{\alpha} \leq \hat{\alpha}$.

◇ Partial exposure: If the opportunistic A plays the mixed strategy, denote $x^* \in (0,1)$ as her equilibrium probability to litigate. B's equilibrium belief upon litigation therefore is α_x^* in condition (1), which in turn determines $e_{B,x}^*$. Since only the opportunistic A settles, the settlement offer $s^* = u_B(\bar{\theta})$, and she is willing to play mixed strategy iff $\pi - u_B(\bar{\theta}) = u_A(\bar{\theta}, e_{B,x}^*)$. This guarantees that the good A won't deviate to offer s^*. By $\alpha_x^* \in (\hat{\alpha}, 1)$ and so $e_{B,x}^* \in (\underline{e}_B, e_B^*(\hat{\theta}))$, we can find such $e_{B,x}^*$ iff $\pi - u_B(\bar{\theta}) \in (u_A(\bar{\theta}, e_B^*(\hat{\theta})), u_A(\bar{\theta}, \underline{e}_B))$. To support this equilibrium, B should reject any deviant offer $\tilde{s} < u_B(\bar{\theta})$ and litigate with $\tilde{e}_B \geq e_{B,x}^*$. In other words, B should put enough weight on the opportunistic A upon receiving $\tilde{s} < u_B(\bar{\theta})$.

To show that both equilibria survive $D1$, it suffices to show that the opportunistic A cannot be deleted in B's off-path beliefs. Since A's equilibrium payoff is $u_A(\theta, e_B)$, depending on A's type and the prevailing e_B for each equilibrium, upon a deviation offer, B's response of rejection and litigation with the equilibrium efforts level makes both types of A indifferent from deviation or not. And by $u_A(\bar{\theta}, e_B) < u_A(\underline{\theta}, e_B)$, whenever B's acceptance of a deviant offer makes the good A weakly better-off by deviating, the opportunistic A strictly prefers that deviation. Summing up, $D1$ cannot rule out the opportunistic type.

For other bargaining outcomes:

◇ No litigation: The minimal offer to settle with both types of A is $u_B(\hat{\theta})$. Let it be an equilibrium payment. To support this equilibrium, let B accept any deviant offers larger than $u_B(\hat{\theta})$ with, say, "passive belief" $\hat{\theta}$. When facing a smaller offer, B should reject it and exert litigation effort $\tilde{e}_B$ such that $u_A(\underline{\theta}, \tilde{e}_B) \leq \pi - u_B(\hat{\theta})$. But by LEMMA

1, $D1$ requires that B's belief upon such offer be the good type for sure, which in turn require B to accept any offer in $(u_B(\underline{\theta}), u_B(\hat{\theta}))$. Therefore no PBE fulfilling $D1$ can implement this outcome. On the other hand, since the passive belief is allowed under divinity, and u_A is decreasing in e_B, no litigation can be implemented by a PBE satisfying divinity if $u_A(\underline{\theta}, e_B^*(\hat{\theta})) \leq \pi - u_B(\hat{\theta})$.

$\diamond$ Only the good A litigates: First consider a full separating equilibrium such that the good A always litigates while the opportunistic A always settles. In this case, the opportunistic A's equilibrium offer is $u_B(\bar{\theta})$, and the good A litigates against an effort $\underline{e}_B$. Neither type will deviate to play the other's equilibrium strategy when $u_A(\underline{\theta}, \underline{e}_B) \geq \pi - u_B(\bar{\theta}) \geq u_B(\bar{\theta}, \underline{e}_B)$. No inventor would offer higher than $u_B(\bar{\theta})$ to settle the case; for a deviant offer $\tilde{s} < u_B(\bar{\theta})$, the equilibrium can only be supported by B's rejecting $\tilde{s}$ and litigating with $\tilde{e}_B \geq \underline{e}_B$. Since A can be sure to face the minimal effort $\underline{e}_B$ by proposing the good A's offer (it could be an empty offer), no patent-holder has incentives to deviate to any other offers strictly smaller than $u_B(\underline{\theta})$.

Consider a deviant offer $\tilde{s} \in [u_B(\underline{\theta}), u_B(\bar{\theta}))$. To reject this offer, B should put enough weight on the opportunistic type, i.e., $\tilde{\theta}$ so high that $\tilde{s} < u_B(\tilde{\theta})$. We show that for $\tilde{s}$ small enough, $D1$ would require $Pr(\underline{\theta}|\tilde{s}) = 1$ and so this outcome cannot be supported as an equilibrium outcome. Relaxing the requirement to divinity, this outcome is possible only when $\hat{\alpha}$ small enough. Denote $(\tilde{\phi}, \tilde{e}_B)$ as B's optimal response to $\tilde{s}$, which is rationalized by belief $\tilde{\alpha}$.

If $\tilde{s} \in [\pi - u_A(\underline{\theta}, \underline{e}_B), u_B(\bar{\theta}))$, B's response $\tilde{\phi} = 1$ makes the opportunistic A strictly better off but not the good A, relative to their equilibrium payoffs; $D1$ and divinity cannot constrain $\tilde{\theta}$. For $\tilde{s} \in [u_B(\underline{\theta}), \pi - u_A(\underline{\theta}, \underline{e}_B))$, (i) if $\tilde{\phi} = 1$, both types of A strictly prefer $\tilde{s}$ than their equilibrium strategy; (ii) if $\tilde{\phi} = 0$ and $\pi - u_B(\bar{\theta}) > u_A(\bar{\theta}, \underline{e}_B)$, whatever $\tilde{e}_B$, this response cannot make the good (opportunistic) A strictly (weakly, respectively) better off; and (iii) if $\tilde{\phi} \in (0, 1)$, then for B to take mixed strategy response, $\tilde{s} = u_B(\tilde{\theta})$ and $\tilde{e}_B = e_B^*(\tilde{\theta})$. The opportunistic A weakly prefers to deviate if

$$\tilde{\phi}(\pi - \tilde{s}) + (1 - \tilde{\phi}) u_A(\bar{\theta}, \tilde{e}_B) \geq \pi - u_B(\bar{\theta}) \Rightarrow \tilde{\phi} \geq \bar{\phi} \equiv \frac{\pi - u_B(\bar{\theta}) - u_A(\bar{\theta}, \tilde{e}_B)}{\pi - u_B(\tilde{\theta}) - u_A(\bar{\theta}, \tilde{e}_B)};$$

and the good A strictly prefers to deviate if

$$\begin{aligned} & \tilde{\phi}(\pi - \tilde{s}) + (1 - \tilde{\phi}) u_A(\underline{\theta}, \tilde{e}_B) > u_A(\underline{\theta}, \underline{e}_B) \\ \Rightarrow \quad & \pi - u_B(\tilde{\theta}) > u_A(\underline{\theta}, \tilde{e}_B) \ \text{ and } \ \tilde{\phi} > \underline{\phi} \equiv \frac{u_A(\underline{\theta}, \underline{e}_B) - u_A(\underline{\theta}, \tilde{e}_B)}{\pi - u_B(\tilde{\theta}) - u_A(\underline{\theta}, \tilde{e}_B)}. \end{aligned}$$

$D1$ and divinity have no bite for those $\tilde{s}$ such that $\pi - u_B(\tilde{\theta}) \leq u_A(\underline{\theta}, \tilde{e}_B)$. But this

won't be the case for all $\tilde{\theta}$, for $\pi > u_A(\underline{\theta}, \underline{e}_B) + u_B(\underline{\theta})$ as $\tilde{\theta} \to \underline{\theta}$ (as $\tilde{s} \to u_B(\underline{\theta})$). Define $\tilde{S} \equiv \{\tilde{s} : u_A(\underline{\theta}, \tilde{e}_B) + u_B(\tilde{\theta}) < \pi, \bar{\phi} > \underline{\phi}\}$. $\tilde{S} \neq \emptyset$ since, as $\tilde{s} \to u_B(\underline{\theta})$,

$$\bar{\phi} \to \frac{\pi - u_B(\bar{\theta}) - u_A(\bar{\theta}, \underline{e}_B)}{\pi - u_B(\underline{\theta}) - u_A(\bar{\theta}, \underline{e}_B)} > 0, \quad \text{but} \quad \underline{\phi} \to \frac{u_A(\underline{\theta}, \underline{e}_B) - u_A(\underline{\theta}, \underline{e}_B)}{\pi - u_B(\underline{\theta}) - u_A(\underline{\theta}, \underline{e}_B)} = 0.$$

For all $\tilde{s} \in \tilde{S}$, the set of B's strictly mixed strategy best responses that makes the good A strictly prefers to deviate is strictly larger than the set that makes the opportunistic A weakly prefers to deviate. Therefore, for any $s' \in S' \equiv \tilde{S} \cap [u_B(\underline{\theta}), \pi - u_A(\underline{\theta}, \underline{e}_B))$, $D1$ requires B to hold belief $\theta' = \underline{\theta}$, and divinity requires a belief $\theta' \leq \hat{\theta}$. Imposing $D1$ then eliminates this full separating equilibrium, as B should accept the offer $u_B(\underline{\theta})$. And divinity will bust the equilibrium when $\hat{\alpha}$ is so large, and $\hat{\theta}$ so small that $u_B(\hat{\theta}) \leq s'$ for some $s' \in S'$, since B needs to reject s' with some θ' such that $u_B(\theta') > s'$.

Lastly, suppose $\pi - u_B(\bar{\theta}) = u_A(\bar{\theta}, \underline{e}_B)$. In this case $D1$ and divinity have no bite for (i) when $\tilde{s} = u_B(\underline{\theta})$, B's response $\tilde{\phi} = 0$ and $\tilde{e}_B = \underline{e}_B$ makes both types of A indifferent between deviation or not; and (ii) when $\tilde{s} \in (u_B(\underline{\theta}), \pi - u_A(\underline{\theta}, \underline{e}_B))$,

$$\bar{\phi} = \frac{u_A(\bar{\theta}, \underline{e}_B) - u_A(\bar{\theta}, \tilde{e}_B)}{\pi - u_B(\hat{\theta}) - u_A(\bar{\theta}, \tilde{e}_B)} = \frac{\bar{\theta}(\tilde{e}_B - \underline{e}_B)\pi}{\bar{\theta}\tilde{e}_B\pi - u_B(\tilde{\theta})} < \underline{\phi} = \frac{\underline{\theta}(\tilde{e}_B - \underline{e}_B)\pi}{\underline{\theta}\tilde{e}_B\pi - u_B(\tilde{\theta})},$$

even when $\pi - u_B(\hat{\theta}) - u_A(\underline{\theta}, \tilde{e}_B) > 0$.

$\diamond$ The good A plays mixed strategies: Lastly, if the good A plays the mixed strategy, denote y^* as her equilibrium probability to settle. B's belief upon settlement then is α_y^*, with $\theta_y^* = \alpha_y^* \underline{\theta} + (1 - \alpha_y^*)\bar{\theta}$, and the equilibrium settlement offer $s^* = u_B(\theta_y^*)$, such that

$$u_A(\underline{\theta}, \underline{e}_B) = \pi - u_B(\theta_y^*) \quad \text{and} \quad \alpha_y^* = \frac{\hat{\alpha} y^*}{\hat{\alpha} y^* + 1 - \hat{\alpha}}.$$

Since only the good A litigates, the equilibrium litigation effort is $\underline{e}_B$. The good A is willing to play a mixed strategy iff $u_A(\underline{\theta}, \underline{e}_B) = \pi - u_B(\theta_y^*)$, which leaves the opportunistic A no incentives to deviate and litigate. Since $\alpha_y^* \in (0, \hat{\alpha})$ and so $u_B(\theta_y^*) \in (u_B(\hat{\theta}), u_B(\bar{\theta}))$, this equilibrium requires $u_A(\underline{\theta}, \underline{e}_B) \in (\pi - u_B(\bar{\theta}), \pi - u_B(\hat{\theta}))$. Note that any deviant offer leading to litigation won't disturb this equilibrium, for the inventor's equilibrium payoff is $\pi - u_B(\theta_y^*) = u_A(\underline{\theta}, \underline{e}_B) > u_A(\bar{\theta}, \underline{e}_B)$. To support this equilibrium, we need only to check that there is belief satisfying divinity and inducing B's rejection of a deviant offer $\tilde{s} \in [u_B(\bar{\theta}), u_B(\theta_y^*))$. Since $\alpha_y^* < \hat{\alpha}$ and so $u_B(\hat{\theta}) < u_B(\theta_y^*)$, (i) for $\tilde{s} \in [u_B(\underline{\theta}), u_B(\hat{\theta}))$, whether divinity can trim B's off-path belief, upon deviation we can use the 'passive belief' $\hat{\theta}$ to justify B's rejection; and (ii) for $\tilde{s} \in [u_B(\hat{\theta}), u_B(\theta_y^*))$, it

can be rejected only with belief $\tilde{\theta}$ such that $u_B(\tilde{\theta}) > \tilde{s} \geq u_B(\hat{\theta})$, and so to have $\tilde{\theta} > \hat{\theta}$ the weight on the opportunistic A should not be constrained by divinity. B's accepting $\tilde{s}$ makes both types of A strictly better off; his rejection, together with litigation effort strictly higher than $\underline{e}_B$ makes A worse off. But if B plays a mixed strategy composed of $\tilde{\phi} \in (0,1)$ and $\tilde{e}_B$, since A's equilibrium payoff doesn't not depend on her type, and

$$\tilde{\phi}(\pi - \tilde{s}) + (1 - \tilde{\phi})u_A(\underline{\theta}, \tilde{e}_B) > \tilde{\phi}(\pi - \tilde{s}) + (1 - \tilde{\phi})u_A(\bar{\theta}, \tilde{e}_B),$$

whenever the opportunistic A weakly prefers to deviate, the good A strictly prefers to do so. For this range of $\tilde{s}$, divinity then requires off-path belief $\tilde{\theta} \leq \hat{\theta}$, and so this equilibrium cannot survive $wD1$. *Q.E.D.*

□ Proposition 4

Proof. The necessary and sufficient conditions come directly from $d[e_P + e_B^*(\hat{\theta})]/de_P \leq 0$ and $dC(e_P)/de_P > 0$. The sufficient condition of no lower examination standard is obtained by setting $e_P = 0$ in condition (3), and the necessary condition of no larger cost is obtained by inserting $(de_B^*/d\hat{\alpha})(\partial\hat{\alpha}/\partial e_P) \leq -1$ into $dC(e_P)/de_P > 0$. *Q.E.D.*

□ Proposition 6

Proof. In a separating equilibrium where only the good A settles, along the equilibrium path the settlement payment $\underline{s}$ is determined by B's indifference between accepting the offer or litigating against the good A. Note that upon settlement, B receives $\underline{s}$ only when the application survives subsequent public enforcement e_P^n. And the opportunistic A faces private challenge efforts $\bar{e}_B$, and public examination e_P^c if survives the challenge. Condition (5) comes from that neither type of A is willing to deviate to mimic the other type. That is, the good A prefers paying $\underline{s}$ than encountering two stages of enforcement, $(1 - \underline{\theta}e_P^n)(\pi - \underline{s}) \geq (1 - \underline{\theta}\bar{e}_B)(1 - \underline{\theta}e_P^c)\pi$; and the opportunistic A prefers examination than settlement, $(1 - \bar{\theta}\bar{e}_B)(1 - \bar{\theta}e_P^c)\pi \geq (1 - \bar{\theta}e_P^n)(\pi - \underline{s})$. To support this equilibrium, B accepts any deviant offer $s' > \underline{s}$, and rejects any $s' < \underline{s}$ whiling litigating with efforts $\bar{e}_B$. *Q.E.D.*

B Alternative settings

This appendix shows that the main results we obtained are robust to alternative settings where (i) the potential challenger B makes the settlement offer; or (ii) A's possible

types are continuous.

□ **When B makes the offer:** Assign the whole bargaining power to B in the two-type case. Given belief $\hat{\alpha}$, and so average invalidity $\hat{\theta}$, if B decides not to settle at all, his expected payoff from litigation is $u_B(\hat{\theta})$. If he wants to settle only with the opportunistic A, the settlement offer (the payoff he promises to A) is $u_A(\bar{\theta}, \underline{e}_B)$, and he will exert effort $\underline{e}_B$ against the good A (recall that this effort cannot be part of the settlement agreement). His payoff under this "partial settlement" policy is $\hat{\alpha}u_B(\underline{\theta}) + (1-\hat{\alpha})[\pi - u_A(\bar{\theta}, \underline{e}_B)]$.

To fully settle the case A's willingness to accept B's offer depends on the e_B at the off-path event of litigation, and a higher e_B pushes down the settlement offer. But next proposition shows that only $\underline{e}_B$ fulfills the criterion $D1$.[23] By offering $u_A(\underline{\theta}, \underline{e}_B)$, B's payoff from fully settlement is $\pi - u_A(\underline{\theta}, \underline{e}_B)$. Define the following terms:

$$\bar{\alpha}_1: \quad \pi - u_A(\underline{\theta}, \underline{e}_B) \equiv \bar{\alpha}_1 u_B(\underline{\theta}) + (1-\bar{\alpha}_1)[\pi - u_A(\bar{\theta}, \underline{e}_B)] \;\Rightarrow\; \bar{\alpha}_1 \equiv \frac{(\bar{\theta}-\underline{\theta})\underline{e}_B\pi}{\bar{\theta}\underline{e}_B\pi - u_B(\underline{\theta})},$$

$$\bar{\alpha}_2: \quad u_B(\bar{\alpha}_2\underline{\theta} + (1-\bar{\alpha}_2)\bar{\theta}) \equiv \pi - u_A(\underline{\theta}, \underline{e}_B), \text{ and}$$

$$\bar{\alpha}_3: \quad u_B(\bar{\alpha}_3\underline{\theta} + (1-\bar{\alpha}_3)\bar{\theta}) \equiv \bar{\alpha}_3 u_B(\underline{\theta}) + (1-\bar{\alpha}_3)[\pi - u_A(\bar{\theta}, \underline{e}_B)], \text{ s.t. } \bar{\alpha}_3 < 1.$$

$\bar{\alpha}_1$ is the cutoff level where B is indifferent between full settlement and settling only with the opportunistic inventor (partial settlement). By the same token, $\bar{\alpha}_2$ is the cutoff where B is indifferent between no settlement at all and full settlement; and $\bar{\alpha}_3$ the cutoff for indifference between no settlement and partial settlement. Note that $\bar{\alpha}_1 \in (0,1)$ is always well-defined, but there not may exist $\bar{\alpha}_2$ and $\bar{\alpha}_3$ in the open interval $(0,1)$.

PROPOSITION 8. *(Bargaining equilibria when B makes the offer) Let B make the settlement offer. Suppose that A agrees to settle whenever she is indifferent between settlement or not, the offer to fully settle the case in a PBE surviving $D1$ is $u_A(\underline{\theta}, \underline{e}_B)$. In this case, the weak patent is fully exposed to private enforcement only when $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$, and (i) $\hat{\alpha} < \bar{\alpha}_2$, in the case of $\bar{\alpha}_1 \leq \bar{\alpha}_2$; or (ii) $\hat{\alpha} < \bar{\alpha}_3$, in the case of $\bar{\alpha}_1 > \bar{\alpha}_2$. Otherwise, either there is no litigation or only the good A litigates.*

Suppose that A may also respond to B's offer in mixed strategies, then B's payoff is strictly higher when the weak patent is only partially exposed to private enforcement than when full exposure. When $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$ and $\hat{\alpha}$ small enough so that

[23] However, the general pattern of bargaining outcomes is not affected by this selection.

full litigation is optimal in the previous case, it is optimal for B to make a settlement offer $u_A(\bar{\theta}, e_B^*(\theta_z))$ and exert litigation efforts $e_B^*(\theta_z)$ such that the opportunistic A will litigate with probability $z \in (0,1)$ and the good A will always litigate, where $\theta_z = \alpha_z \underline{\theta} + (1-\alpha_z)\bar{\theta}$ and $\alpha_z \equiv \hat{\alpha}/[\hat{\alpha} + (1-\hat{\alpha})z] \in (\hat{\alpha}, 1)$. B's payoff is

$$\max_{\alpha_z} U_z = \frac{\hat{\alpha}}{\alpha_z} u_B(\theta_z) + (1 - \frac{\hat{\alpha}}{\alpha_z})[\pi - u_A(\bar{\theta}, e_B^*(\theta_z))].$$

Proof. Suppose that A will agree to settle upon indifference. To fully settle the case, B needs to offer a payoff $u_A(\underline{\theta}, e)$, where $e \in [\underline{e}_B, \bar{e}_B]$ is determined by B's off-path belief should A reject the offer. The lowest offer, $u_A(\underline{\theta}, \bar{e}_B)$, is supported by the belief that the rejection must come from the opportunistic A. This, however, doesn't satisfy $D1$, according to LEMMA 1. This lemma also shows that the only off-path belief surviving $D1$ is that such rejection must be from the good type; and so the offer could be supported by a PBE with $D1$ is $u_A(\underline{\theta}, \underline{e}_B)$. By comparing B's payoffs from different settlement policies, we get the range of $\hat{\alpha}$ such that B will not settle at all.

Suppose that A can respond to B's offer with mixed strategies. First note that it won't be in B's interests to let the good A play a mixed strategy. In that case, B offers a payoff $u_A(\underline{\theta}, \underline{e}_B)$ so that the good A is indifferent between settlement and litigation; and since the opportunistic A always settles, the litigation effort is $\underline{e}_B$. The good A's probability of acceptance will only change the belief upon settlement, but neither the settlement offer nor the litigation effort. By $\pi - u_A(\underline{\theta}, \underline{e}_B) > u_B(\underline{\theta})$, B's payoff is increasing in the probability of the good A's settlement; B can increase his offer by a very small amount to guarantee full settlement.

Now, suppose that opportunistic A adopts mixed-strategy responses. Given $\hat{\alpha}$, if she litigates with probability $z \in (0,1)$ upon indifference, then B's belief upon litigation becomes $\alpha_z \equiv \hat{\alpha}/[\hat{\alpha} + (1-\hat{\alpha})z] \in (\hat{\alpha}, 1)$, and litigation efforts $e_B^*(\theta_z) \in (\underline{e}_B, e_B^*(\hat{\theta}))$. As z increases, α_z decreases and $e_B^*(\theta_z)$ increases. For the opportunistic A to be indifferent, B offers a settlement payoff $u_A(\bar{\theta}, e_B^*(\theta_z))$. By doing so, B's payoff is

$$\begin{aligned} U_z &= \hat{\alpha}[\underline{\theta} e_B^*(\theta_z) b - c(e_B^*(\theta_z))] \\ &\quad + (1-\hat{\alpha})\Big\{ z[\bar{\theta} e_B^*(\theta_z) b - c(e_B^*(\theta_z))] + (1-z)[\pi - u_A(\bar{\theta}, e_B^*(\theta_z))] \Big\} \\ &= [\hat{\alpha} + (1-\hat{\alpha})z] u_B(\theta_z) + (1-\hat{\alpha})(1-z)[\pi - u_A(\bar{\theta}, e_B^*(\theta_z))] \\ &= \frac{\hat{\alpha}}{\alpha_z} u_B(\theta_z) + (1 - \frac{\hat{\alpha}}{\alpha_z})[\pi - u_A(\bar{\theta}, e_B^*(\theta_z))]. \end{aligned}$$

B can obtain a payoff $U_z(\alpha_z)$, with any $\alpha_z \in (\hat{\alpha}, 1)$, when opportunistic A sets $z = [\hat{\alpha}(1-\alpha_z)]/[(1-\hat{\alpha})\alpha_z]$.

Note that as $\alpha_z \to \hat{\alpha}$, $U_z \to u_B(\hat{\theta})$, B's payoff under no settlement; and

$$\begin{aligned}\left.\frac{du_B(\theta_z)}{d\alpha_z}\right|_{\alpha_z=\hat{\alpha}} &= \frac{1}{\hat{\alpha}}[\pi - u_A(\bar{\theta}, e_B^*(\hat{\theta})) - u_B(\hat{\theta})] + \frac{du_B(\hat{\theta})}{d\hat{\alpha}} + (1-\frac{\hat{\alpha}}{\hat{\alpha}})\frac{du_A(\bar{\theta}, e_B^*(\hat{\theta}))}{de_B}\frac{\partial e_B^*(\hat{\theta})}{\partial\hat{\alpha}} \\ &= \frac{1}{\hat{\alpha}}\left[\pi - u_A(\bar{\theta}, e_B^*(\hat{\theta})) - u_B(\hat{\theta}) - (\bar{\theta}-\underline{\theta})e_B^*(\hat{\theta})b\right] \\ &> \frac{1}{\hat{\alpha}}\bar{\theta}(\pi - b)e_B^*(\hat{\theta}).\end{aligned}$$

Full litigation is strictly dominated when A plays mixed strategies. This implies that, when $\hat{\alpha}$ is small enough so that B doesn't want to settle at all in case where A always settles upon indifference, it is optimal for B to obtain a payoff U_z. On the other hand, when $\hat{\alpha} \to 1$, the feasible set of α_z, $(\hat{\alpha}, 1)$ shrinks, and $U_z \to u_B(\underline{\theta})$, which is strictly smaller than $\pi - u_A(\underline{\theta}, \underline{e}_B)$, the payoff from full litigation. Therefore for $\hat{\alpha}$ large enough, it won't be optimal for B to induce mixed-strategy response from A. *Q.E.D.*

Comparing this proposition with PROPOSITION 2, the same condition, $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$, applies for the weak patent to be subject to private enforcement. However, since $u_B(\hat{\theta})$ is increasing in $\hat{\theta}$ and so decreasing in $\hat{\alpha}$, a higher patent quality makes settlement more attractive to B. Unlike the case where A makes the offer, in this case the opportunistic A is fully exposed to private enforcement only when the patent quality is low enough. This is the major difference between the two distributions of bargaining power.

But, in fact, in this case the full and partial exposure regimes take place for the same range of $\hat{\alpha}$. Different regimes ensue depending on whether A is allowed to play mixed strategies, and B's payoff improves when the opportunistic A can be induced to play mixed strategies in a proper manner, and so only litigates with some probability.

Consider the impact of e_P on different regimes. Under full exposure, there is no settlement, and B's litigation effort is $e_B^*(\hat{\theta})$. The crowding out effect of public enforcement thus is robust to the distribution of bargaining power. And under partial exposure, we show in the following proposition that a positive relationship between public and private enforcement still obtains with some additional conditions.

PROPOSITION 9. *(Partial exposure when B makes the offer) When B makes the offer, the weak patent may encounter a private challenge only when* $u_A(\bar{\theta}, \underline{e}_B) > \pi - u_B(\bar{\theta})$, *and at the full exposure regime a higher* e_P *reduces B's litigation efforts.*

Under the partial exposure, if B's cost function $c''' \geq 0$ *and* $\hat{\alpha}$ *small enough, then B's litigation efforts is independent of* e_P *and the opportunistic A's litigation probability is increasing in* e_P.

Proof. When B makes the offer and the opportunistic A litigates with probability $z \in (0,1)$ upon indifference, by the proof of PROPOSITION 8 for $\hat{\alpha}$ smaller than $\bar{\alpha}_2$ or $\bar{\alpha}_3$, depending on $\bar{\alpha} \gtrless \bar{\alpha}_2$, it is optimal for B to induce the mixed-strategy response from the opportunistic A and obtain a payoff U_z for some z.

Given such $\hat{\alpha}$, denote $\alpha_z^* \in (\hat{\alpha}, 1)$ as the optimal belief upon litigation (derived from the optimal z^*), and $\theta_z^* = \alpha_z^* \underline{\theta} + (1-\alpha_z^*)\bar{\theta}$. B's optimal payoff is

$$\begin{aligned} U_z(\theta_z^*) &= \frac{\hat{\alpha}}{\alpha_z^*} u_B(\theta_z^*) + (1 - \frac{\hat{\alpha}}{\alpha_z^*})[\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*))] \\ &= \pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - \frac{\hat{\alpha}}{\alpha_z^*}\Big[\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - u_B(\theta_z^*)\Big]. \end{aligned}$$

When $c''' \geq 0$, for all $\hat{\alpha}$, U_z is strictly convex in α_z:

$$\begin{aligned} FOC: \ \frac{\partial U_z}{\partial \alpha_z} &= \bar{\theta}\pi \frac{\partial e_B^*(\theta_z)}{\partial \alpha_z} + \frac{\hat{\alpha}}{\alpha_z^2}[\bar{\theta}\pi e_B^*(\theta_z) - u_B(\theta_z)] - \frac{\hat{\alpha}}{\alpha_z}[\bar{\theta}\pi \frac{\partial e_B^*(\theta_z)}{\partial \alpha_z} + (\bar{\theta} - \underline{\theta}) b e_B^*(\theta_z)], \\ SOC: \ \frac{\partial^2 U_z}{\partial \alpha_z^2} &= -\frac{2\hat{\alpha}}{\alpha_z^3}\left[\bar{\theta} e_B^*(\theta_z)(\pi - \alpha_z b) + c(e_B^*(\theta_z)) + (\bar{\theta} - \underline{\theta})\alpha_z b \frac{\bar{\theta}(\pi - \alpha_z b) + \underline{\theta}\alpha_z b}{c'(e_B^*(\theta_z))}\right] \\ &\quad + \bar{\theta}\pi(1 - \frac{\hat{\alpha}}{\alpha_z}) \frac{\partial^2 e_B^*(\theta_z)}{\partial \alpha_z^2} < 0, \end{aligned}$$

where

$$\frac{\partial^2 e_B^*(\theta_z)}{\partial \alpha_z^2} = \frac{c'''}{(c'')^2}(\bar{\theta} - \underline{\theta}) b \frac{\partial e_B^*(\theta_z)}{\partial \alpha_z} \leq 0.$$

Together with $\partial U_z / \partial \alpha_z > 0$ as $\alpha_z \to \hat{\alpha}$ and $U_z \to \hat{\alpha} u_B(\underline{\theta}) + (1-\hat{\alpha})[\pi - u_A(\bar{\theta}, \underline{e}_B)]$ as $\alpha_z \to 1$, the generalized program $\max_{\alpha_z} U_z$ has a unique solution over $\alpha_z \in (\hat{\alpha}, 1]$. If $\partial U_z / \partial \alpha_z < 0$ as $\alpha_z \to 1$, then the optimal $\alpha_z^* \in (\hat{\alpha}, 1)$; and if $\partial U_z / \partial \alpha_z \geq 0$ as $\alpha_z \to 1$, then we have a corner solution and B should fully settle with the opportunistic A. In the former case, as $\alpha_z \to 1$, the first-order condition,

$$\left.\frac{\partial U_z}{\partial \alpha_z}\right|_{\alpha_z \to 1} = \bar{\theta}\pi \left.\frac{\partial e_B^*(\theta_z)}{\partial \alpha_z}\right|_{\alpha_z \to 1} + \hat{\alpha}\left[\bar{\theta}\pi \underline{e}_B - u_B(\underline{\theta}) - \bar{\theta}\pi \left.\frac{\partial e_B^*(\theta_z)}{\partial \alpha_z}\right|_{\alpha_z \to 1} + (\bar{\theta} - \underline{\theta}) b e_B^*(\theta_z)\right],$$

becomes strictly negative for $\hat{\alpha}$ small enough, i.e., we must have an interior solution.

Suppose that $\hat{\alpha}$ is so small that the optimal $\alpha_z^* \in (\hat{\alpha}, 1)$. Considering a small increase in the patent quality $\hat{\alpha}' > \hat{\alpha}$, we show that the same α_z^* remains optimal when $\hat{\alpha}'$ is close enough to $\hat{\alpha}$. Let $\hat{\alpha}'$ be close enough to $\hat{\alpha}$ so that $\alpha_z^* \in (\hat{\alpha}', 1)$. We want to

show that $\forall \alpha' \in (\hat{\alpha}', 1)$ and $\alpha' \neq \alpha_z^*$, with $\theta' = \alpha' \underline{\theta} + (1-\alpha')\bar{\theta}$,

$$\begin{aligned} & \pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - \frac{\hat{\alpha}'}{\alpha_z^*}\Big[\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - u_B(\theta_z^*)\Big] \\ > \quad & \pi - u_A(\bar{\theta}, e_B^*(\theta')) - \frac{\hat{\alpha}'}{\alpha'}\Big[\pi - u_A(\bar{\theta}, e_B^*(\theta')) - u_B(\theta')\Big], \\ \Rightarrow \quad & u_A(\bar{\theta}, e_B^*(\theta')) - u_A(\bar{\theta}, e_B^*(\theta_z^*)) > \hat{\alpha}'\left\{\frac{\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - u_B(\theta_z^*)}{\alpha_z^*} - \frac{\pi - u_A(\bar{\theta}, e_B^*(\theta')) - u_B(\theta')}{\alpha'}\right\}. \end{aligned}$$

By the definition and uniqueness of α_z^*, since α' is also available under $\hat{\alpha}$ (for $(\hat{\alpha}', 1) \subset (\hat{\alpha}, 1)$),

$$\begin{aligned} & \pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - \frac{\hat{\alpha}}{\alpha_z^*}\Big[\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - u_B(\theta_z^*)\Big] \\ > \quad & \pi - u_A(\bar{\theta}, e_B^*(\theta')) - \frac{\hat{\alpha}}{\alpha'}\Big[\pi - u_A(\bar{\theta}, e_B^*(\theta')) - u_B(\theta')\Big] \\ \Rightarrow \quad & u_A(\bar{\theta}, e_B^*(\theta')) - u_A(\bar{\theta}, e_B^*(\theta_z^*)) > \hat{\alpha}\left\{\frac{\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - u_B(\theta_z^*)}{\alpha_z^*} - \frac{\pi - u_A(\bar{\theta}, e_B^*(\theta')) - u_B(\theta')}{\alpha'}\right\}. \end{aligned}$$

Therefore, if $\alpha' < \alpha_z^*$, then $e_B(\theta') > e_B(\theta_z^*)$ and so $u_A(\bar{\theta}, e_B^*(\theta')) < u_A(\bar{\theta}, e_B^*(\theta_z^*))$, any $\hat{\alpha}' > \hat{\alpha}$ will fulfill our objective. The same is true if $\alpha' > \alpha_z^*$ but

$$\frac{\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - u_B(\theta_z^*)}{\alpha_z^*} \leq \frac{\pi - u_A(\bar{\theta}, e_B^*(\theta')) - u_B(\theta')}{\alpha'}.$$

On the other hand, if $\alpha' > \alpha_z^*$ and

$$\frac{\pi - u_A(\bar{\theta}, e_B^*(\theta_z^*)) - u_B(\theta_z^*)}{\alpha_z^*} > \frac{\pi - u_A(\bar{\theta}, e_B^*(\theta')) - u_B(\theta')}{\alpha'},$$

a $\hat{\alpha}'$ close enough to $\hat{\alpha}$ guarantees the optimality of α_z^* under $\hat{\alpha}'$. *Q.E.D.*

□ **Continuous types:** Now, let A keep the bargaining power, but assume continuous types $\theta \in [0, 1]$. Let *ex ante*, i.e., before the examination process begins, CDF be $F(\cdot)$ and *pdf* be $f(\cdot)$, with $f(\theta) > 0$ for all $\theta \in [0, 1]$. Again denote $\theta^0 \equiv \int_0^1 \theta dF$ as the *ex ante* expectation value of θ. A higher θ^0 implies a lower quality.

When all types of inventors file patent applications, under the post-grant challenge system and patent office efforts e_P, the probability to eliminate the application is $\int_0^1 \theta e_P dF = \theta^0 e_P$. Upon issuance, the distribution of θ is updated to

$$\hat{F}(\theta) \equiv \frac{1}{1-\theta^0 e_P}\int_0^\theta (1-\theta' e_P) dF \quad \text{and} \quad \hat{f}(\theta) \equiv \frac{1-\theta e_P}{1-\theta^0 e_P} f(\theta);$$

and the post-issuance expectation is

$$\hat{\theta} \equiv \int_0^1 \theta d\hat{F} = \frac{\theta^0 - e_P E(\theta^2)}{1 - e_P \theta^0}.$$

Intuitively, stronger public enforcement reduces $\hat{\theta}$:

$$\frac{\partial \hat{\theta}}{\partial e_P} = \frac{(\theta^0)^2 - E(\theta^2)}{(1 - e_P \theta^0)^2} \leq 0,$$

by Jensen's inequality and the fact that x^2 is a convex function.

To facilitate the presentation, let us define the following terms: given $\tilde{\theta} \in (0, 1)$,

$$\hat{\theta}^+ \equiv E(\theta | \theta \geq \tilde{\theta}, e_P) = \frac{1}{1 - \hat{F}(\tilde{\theta})} \int_{\tilde{\theta}}^1 \theta d\hat{F} \text{ and } \theta^+ \equiv E(\theta | \theta \geq \tilde{\theta}, e_P = 0) = \frac{1}{1 - F(\tilde{\theta})} \int_{\tilde{\theta}}^1 \theta dF.$$

$\hat{\theta}^+$ is the post-issuance expectation, conditional on θ greater than a threshold $\tilde{\theta}$; and θ^+ is the conditional mean at the *ex ante* stage, or, equivalently, when $e_P = 0$. By the same token, we define $\hat{\theta}^-$ and θ^- as the conditional expectations when $\theta \leq \tilde{\theta}$:

$$\hat{\theta}^- \equiv E(\theta | \theta \leq \tilde{\theta}, e_P) = \frac{1}{\hat{F}(\tilde{\theta})} \int_0^{\tilde{\theta}} \theta d\hat{F} \text{ and } \theta^- \equiv E(\theta | \theta \leq \tilde{\theta}, e_P = 0) = \frac{1}{F(\tilde{\theta})} \int_0^{\tilde{\theta}} \theta dF.$$

Maintain the assumption that B's litigation effort e_B cannot be part of the settlement agreement. Denote again $u_B(E(\theta|\mathcal{L}))$ as B's expected payoff when challenging a patent with expected "case quality" $E(\theta|\mathcal{L})$. Upon bargaining breakdown, the optimal litigation effort e_B^* also depends on $E(\theta|\mathcal{L})$, and is determined by the first-order condition $E(\theta|\mathcal{L})b \equiv c'(e_B^*)$. Given e_B^*, a patentee with of type θ has a expected payoff from litigation $(1 - \theta e_B^*)\pi$. Since $\theta = 0$ is always one of the possible types, $f(0) > 0$, and cannot be eliminated by the patent office, under asymmetric information full settlement cannot be a bargaining outcome. As long as $Pr(\theta > 0) > 0$, B will not accept an agreement under which A keeps the whole monopoly profit π.

For simplicity, consider only pure strategies. The following proposition, In resemblance of PROPOSITION 1, shows that a settled patent dispute involves weak patents, i.e., those with high values of θ.

PROPOSITION 10. *(Case selection under continuous types) Suppose that both private players use pure strategies. Whether A or B makes the settlement offer, there exists* $\tilde{\theta} \in (0, 1]$ *such that a patent-holder litigates when having types* $\theta' < \tilde{\theta}$*, and settles when having types* $\theta'' > \tilde{\theta}$*.*

Proof. Since only pure strategies are allowed, there is only one equilibrium settlement payment s (from A to B). Without loss of generality, let $s = 0$ if no agreement is ever reached. A bargaining outcome consists of two elements: the equilibrium settlement offer s and B's litigation effort e_B^* in case of bargaining breakdown. A's payoffs from settlement and litigation are $\pi - s$ and $(1-\theta e_B^*)\pi$, respectively. The cut-off rule follows from the fact that the former is constant while the latter is decreasing in θ. *Q.E.D.*

By this proposition, B's equilibrium litigation effort is determined in accordance with the expectation $E(\theta|\mathcal{L}) = \hat{\theta}^-$. Let $\bar{\theta}_A$ be the equilibrium cutoffs. We first derive a sufficient condition under which PBEs exist, then consider the impact of a marginal change in e_P and the possibility of a positive relationship between public and private enforcement.

PROPOSITION 11. *(Bargaining equilibrium with continuous types) Consider continuous types and let A make the settlement offer. If $u_B(1) < e_B^*(\hat{\theta})\pi$, there is no PBE where no types of A settle.*

Any $\bar{\theta}_A \in (0,1)$ is an equilibrium cutoff of a PBE if it satisfies

$$\bar{\theta}_A e_B^*(\hat{\theta}^-)\pi \geq u_B(\hat{\theta}^+) \equiv \max_{e_B} \hat{\theta}^+ e_B b - c(e_B). \tag{7}$$

A sufficient condition for the existence of an equilibrium cutoff $\bar{\theta}_A \in (0,1)$ is

$$e_B^*\left(\frac{\theta^0 - E(\theta^2)}{1-\theta^0}\right)\pi > u_B(1) = \bar{e}_B b - c(\bar{e}_B), \tag{8}$$

where $\bar{e}_B = e_B^*(1) \leq 1$ is the maximal possible litigation effort, and $E(\theta^2)$ is evaluated at the ex ante distribution.

Proof. Firs, consider full litigation as the equilibrium outcome. The equilibrium litigation effort is $e_B^*(\hat{\theta})$, and equilibrium payoff for a patent-holder with type θ is $[1-\theta e_B^*(\hat{\theta})]\pi$, decreasing in θ. To support this equilibrium, B should reject any positive settlement offer with appropriate off-path beliefs. However, since B will always agree to settle when offered a payment $u_B(1)$ (or plus a small amount in order to break the tie), the patentee with types close to $\theta = 1$ will find it profitable to deviate and settle when $\pi - u_B(1) > [1 - e_B^*(\hat{\theta})]\pi$.

Now, suppose that $\bar{\theta}_A \in (0,1)$ is an equilibrium cutoff, i.e., all $\theta' < \bar{\theta}_A$ litigate while all $\theta'' > \bar{\theta}_A$ settle. Let $\hat{\theta}^-$ and $\hat{\theta}^+$ be the conditional means corresponding to $\bar{\theta}_A$.

The type $\bar{\theta}_A$ must be indifferent between litigation (with a payoff $[1 - \bar{\theta}_A e_B^*(\hat{\theta}^-)]\pi$) and settlement (with a payoff $\pi - s$), otherwise she and adjacent types will move toward

the more profitable strategy and upset the equilibrium. The equilibrium settlement payment is $s = \bar{\theta}_A e_B^*(\hat{\theta}^-)\pi$. But this offer has to be no smaller than B's expected payoff from litigating against $\hat{\theta}^+$ in order to accept the offer. Thus determines condition (7). This equilibrium can be supported by B's off-path responses to accept any deviant offers greater than $\bar{\theta}_A e_B^*(\hat{\theta}^-)\pi$, and reject smaller deviant offers while litigate with efforts at least as strong as the equilibrium litigation level $e_B^*(\hat{\theta}^-)$.

For existence, note that as $\bar{\theta}_A \to 1$, $\hat{\theta}^- \to \hat{\theta}$ and $\hat{\theta}^+ \to 1$. The right-hand side of condition (7) is simply B's maximal possible payoff from litigation: $\max_\theta u_B(\theta) = u_B(1) = \bar{e}_B b - c(\bar{e}_B)$. The left-hand side, as $\bar{\theta}_A \to 1$, approaches to $e_B^*(\hat{\theta})\pi$, where $\hat{\theta}$ is decreasing in e_P. To guarantee the existence for all e_P, condition (8) establishes the existence when $e_P \to 1$. *Q.E.D.*

Given an equilibrium cutoff $\bar{\theta}_A \in (0,1)$, the equilibrium settlement payment and litigation efforts are $\bar{\theta}_A e_B^*(\hat{\theta}^-)\pi$ and $e_B^*(\hat{\theta}^-)$, respectively.

REMARK. (EQUILIBRIUM REFINEMENT) As in a typical signaling game, multiple equilibria may ensue.[24] The intuitive criterion has no bites here.[25] And, different from the two-type case, a more stringent criterion such as $D1$ will eliminate all the PBEs with positive probability of settlement. This is because, for all deviant offers $s' \neq s$, those types $\theta'' > \bar{\theta}_A$ will be eliminated under $D1$ by the type $\bar{\theta}_A$: With the same equilibrium payoff but lower probability to be invalidated for all $e_B > 0$, whenever a type θ'' weakly prefers to deviate and offer s', the type $\bar{\theta}_A$ must strictly prefer to do so. But this implies that the highest possible off-path belief is $\bar{\theta}_A$, which busts the equilibrium since B has no reasonable off-path belief to reject a deviant offer s' between $u_B(\bar{\theta}_A)$ and $u_B(\hat{\theta}^+)$. ■

[24]Indeed, when $\pi >> b$ such that

$$\pi\left[e_B^*(\hat{\theta}^-) + \bar{\theta}_A \frac{\partial e_B^*}{\partial \theta}\Big|_{\hat{\theta}^-} \frac{\partial \hat{\theta}^-}{\partial \bar{\theta}_A}\right] > b e_B^*(\hat{\theta}^+)\frac{\partial \hat{\theta}^+}{\partial \bar{\theta}_A},$$

for any $\bar{\theta}_A$ satisfies condition (7), so does any $\theta > \bar{\theta}_A$.

[25]A PBE here can be supported by off-path strategies such that B accepts any deviant payment s' higher than s, and rejects any smaller payment while exerting litigation efforts no smaller than e_B^*. Both responses can be justified by a belief that this offer comes from an inventor with an average type $\hat{\theta}^+$. Note that for $s' < s$, no type of A can be eliminated by the intuitive criterion: Relative to their equilibrium payoffs, B's acceptance of s' is strictly preferred by those $\theta'' > \bar{\theta}$, and the rejection with a litigation effort higher than e_B^* is strictly preferred by $\theta' \leq \bar{\theta}_A$. For the same reason, when $s' > s$, the intuitive criterion won't be able to eliminate a type $\theta' \leq \bar{\theta}_A$. So even if some types $\theta'' > \bar{\theta}_A$ can be deleted, a belief that a deviant offer comes from those types smaller than $\bar{\theta}_A$, with the resulting average quality $\hat{\theta}^-$, suffices to support B's response.

We now proceed to consider the impact of public enforcement e_P. By $\hat{\theta}$ decreasing in e_P, a higher e_P makes it easier to sustain an equilibrium with no settlement. This corresponds to the "full exposure" regime in the two-type case, and requires that the worst type $\theta = 1$ be willing to mix with all other types and fact an litigation effort $e_B^*(\hat{\theta})$ rather than offering $u_B(1)$ to guarantee settlement. This would happen when e_P is high and so $e_B^*(\hat{\theta})$ is low enough.

Now, consider the effect of a marginal change in e_P. An increasing in e_P changes the distribution function $\hat{F}$ at the private bargaining stage: $\forall \theta < 1$,

$$\frac{\partial \hat{F}(\theta)}{\partial e_P} = \frac{\theta^0 - E(\theta'|\theta' \leq \theta)}{(1 - \theta^0 e_P)^2} F(\theta) > 0.$$

A higher public enforcement effort shifts the distribution toward low values of θ. Presumably, this change may simultaneously move the equilibrium cutoff $\bar{\theta}_A$ and effort e_B^*, with the latter both affected by the distribution and the equilibrium cutoff. This makes it difficult to define the extent of private enforcement. For simplicity, we restrict attention to a particular type of equilibrium adjustment. Similar to the partial exposure regime under the two-type case, we consider when an increase in e_P will raise $\bar{\theta}_A$ but keep e_B^* unchanged. If this holds, then a higher public effort enlarges the set of inventor types under private scrutiny without compromising challenge efforts.

We consider a pair of change de_P and $d\bar{\theta}_A$ that keeps $\hat{\theta}^-$ unchanged, and so the equilibrium effort e_B^* unchanged, and test when this pair of changes still satisfies condition (7). Formally, define $\Lambda \equiv \bar{\theta}_A e_B^* \pi - u_B(\hat{\theta}^+)$. In a PBE, $\Lambda \geq 0$. We consider $(de_P, d\bar{\theta}_A)$ such that

$$\frac{\partial \Lambda}{\partial e_P} de_P + \frac{\partial \Lambda}{\partial \bar{\theta}_A} d\bar{\theta}_A \geq 0 \quad s.t. \quad \frac{\partial \hat{\theta}^-}{\partial e_P} de_P + \frac{\partial \hat{\theta}^-}{\partial \bar{\theta}_A} d\bar{\theta}_A = 0. \tag{9}$$

Proposition 12. *(Public and private enforcement under continuous types) In the continuous-type setting where A makes the offer, a higher e_P makes it more likely to have all types of A involved in litigation. Full exposure occurs under high public enforcement.*

In a PBE with equilibrium cutoff $\bar{\theta}_A \in (0,1)$, a pair $(de_P, d\bar{\theta}_A)$ satisfies condition (9) if

$$\frac{\partial \hat{\theta}^-/\partial e_P}{\partial \hat{\theta}^-/\partial \bar{\theta}_A} \geq \frac{\partial \hat{\theta}^+/\partial e_P}{\partial \hat{\theta}^+/\partial \bar{\theta}_A}. \tag{10}$$

Under ex ante uniform distribution $F(\theta) = \theta$, condition (10) is satisfied when $\bar{\theta}_A$ is small enough.

Proof. Since $\hat{\theta}^-$ and so the equilibrium litigation effort e_B^* are not affected by the changes of e_P and $\bar{\theta}_A$, and by definition, $u_B(\hat{\theta}^+) = \hat{\theta}^+ e_B^*(\hat{\theta}^+)b - c(e_B^*(\hat{\theta}^+))$, we have

$$\frac{\partial \Lambda}{\partial e_P} = -e_B^*(\hat{\theta}^+)b\frac{\partial \hat{\theta}^+}{\partial e_P} \quad \text{and} \quad \frac{\partial \Lambda}{\partial \bar{\theta}_A} = e_B^*(\hat{\theta}^-)\pi - e_B^*(\hat{\theta}^+)b\frac{\partial \hat{\theta}^+}{\partial \bar{\theta}_A}.$$

By inserting the condition that keeps $\hat{\theta}^-$ intact,

$$d\bar{\theta}_A = -\frac{\partial \hat{\theta}^-/\partial e_P}{\partial \hat{\theta}^-/\partial \bar{\theta}_A} de_P,$$

and after a few algebraic manipulation, we get

$$\frac{\partial \Lambda}{\partial e_P}de_P + \frac{\partial \Lambda}{\partial \bar{\theta}_A}d\bar{\theta}_A = \frac{de_P}{\partial \hat{\theta}^-/\partial \bar{\theta}_A}\Big[- e_B^*(\hat{\theta}^-)\pi\frac{\partial \hat{\theta}^-}{\partial e_P} + e_B^*(\hat{\theta}^+)b\big(\frac{\partial \hat{\theta}^+}{\partial \bar{\theta}_A}\frac{\partial \hat{\theta}^-}{\partial e_P} - \frac{\partial \hat{\theta}^+}{\partial e_P}\frac{\partial \hat{\theta}^-}{\partial \bar{\theta}_A}\big)\Big].$$

Since $\frac{\partial \hat{\theta}^-}{\partial \theta_A} > 0 > \frac{\partial \hat{\theta}^-}{\partial e_P}$ (and so $d\bar{\theta}_A$ and de_P should have the same sign), the whole term is guaranteed to be positive if

$$\frac{\partial \hat{\theta}^+}{\partial \bar{\theta}_A}\frac{\partial \hat{\theta}^-}{\partial e_P} - \frac{\partial \hat{\theta}^+}{\partial e_P}\frac{\partial \hat{\theta}^-}{\partial \bar{\theta}_A} \geq 0,$$

or, equivalently, if condition (10) holds.

With *ex ante* uniform distribution, $F(\theta) = \theta$, post-issuance CDF and *pdf* are, respectively,

$$\hat{F}(\theta) = \frac{1}{1-\theta^0 e_P}\int_0^\theta (1-\theta' e_P)d\theta' = \frac{\theta(2-\theta e_P)}{2-e_P} \quad \text{and} \quad \hat{f}(\theta) = \frac{2-2\theta e_P}{2-e_P}.$$

Given a cutoff $\bar{\theta}_A$, the conditional expectations are

$$\hat{\theta}^+ = \frac{2}{2-(1+\bar{\theta}_A)e_P}\big[\frac{1}{2}(1+\bar{\theta}_A) - \frac{e_P}{3}(1+\bar{\theta}_A+\bar{\theta}_A^2)\big] \quad \text{and} \quad \hat{\theta}^- = \frac{2\bar{\theta}_A}{2-\bar{\theta}_A e_P}(\frac{1}{2} - \frac{e_P}{3}\bar{\theta}_A).$$

Therefore,

$$\frac{\partial \hat{\theta}^+}{\partial \bar{\theta}_A} = \frac{2(1-\bar{\theta}_A e_P)[2(1-e_P)+(1-\bar{\theta}_A e_P)]}{3[2-(1+\bar{\theta}_A)e_P]^2}, \quad \frac{\partial \hat{\theta}^+}{\partial e_P} = -\frac{(1-\bar{\theta}_A)^2}{3[2-(1+\bar{\theta}_A)e_P]^2},$$
$$\frac{\partial \hat{\theta}^-}{\partial e_P} = -\frac{\bar{\theta}_A^2}{3(2-\bar{\theta}_A e_P)^2}, \quad \frac{\partial \hat{\theta}^-}{\partial \bar{\theta}_A} = \frac{2(3-\bar{\theta}_A e_P)(1-\bar{\theta}_A e_P)}{3(2-\bar{\theta}_A e_P)^2},$$

and condition (10) requires:

$$\frac{\partial \hat{\theta}^-/\partial e_P}{\partial \hat{\theta}^-/\partial \bar{\theta}_A} = -\frac{\bar{\theta}_A^2}{2(3-\bar{\theta}_A e_P)(1-\bar{\theta}_A e_P)} \geq \frac{\partial \hat{\theta}^+/\partial e_P}{\partial \hat{\theta}^+/\partial \bar{\theta}_A} = -\frac{(1-\bar{\theta}_A)^2}{2(1-\bar{\theta}_A e_P)[2(1-e_P)+(1-\bar{\theta}_A e_P)]},$$
$$\Rightarrow (\frac{1-\bar{\theta}_A}{\bar{\theta}_A})^2 \geq \frac{3-\bar{\theta}_A e_P - 2e_P}{3-\bar{\theta}_A e_P}.$$

$\bar{\theta}_A$ has to be small enough. For instance, it is satisfied for all $\bar{\theta}_A \leq \frac{1}{2}$. *Q.E.D.*

References

[1] Ayres, I. and P. Klemperer, (1999), "Limiting Patentees' Market Power without Reducing Innovation Incentives: The Perverse Benefits of Uncertainty and Non-Injunctive Remedies," *Michigan Law Review*, 97: 985-1033.

[2] Bank, J. and J. Sobel, (1987), "Equilibrium Selection in Signalling Games," *Econometrica,* 55 (3): 647-62.

[3] Bebchuk, L., (1984), "Litigation and Settlement under Imperfect Information," *Rand Journal of Economics,* 15 (3): 404-415.

[4] Caillaud, B. and A. Duchêne, (2005), "Patent Office in Innovation Policy: Nobody's Perfect," working paper.

[5] Cho, I.-K. and D. Kreps, (1987), "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics,* 102 (2): 179-222.

[6] Froeb, L. (1993), "Adverse Selection of Case for Trial," *International Review of Law and Economics,* 13 (3): 317-24.

[7] Federal Trade Commission (2003), *To Promote Innovation: A Proper Balance of Competition and Patnet Law and Policy.*

[8] Kesan, J., (2005), "Why "Bad" Patents Survive in the Market and How Should We Change?-The Private and Social Costs of Patents," working paper.

[9] Langinier, C. and P. Marcoul, "Patents, Search of Prior Art and Revelation of Information," working paper.

[10] Lemley, M., (2001), "Rational Ignorance at the Patent Office," *Northwestern University Law Review,* 95(4): 1495-1529.

[11] Merges, R., (1999), "As Many as Six Impossible Patents Before Breakfast: Property Rights for Business Concepts and Patent System Reform," *Berkeley Technology Law Journal,* 14: 577-615.

[12] Meurer, M., (1989), "The Settlement of Patent Litigation," *RAND Journal of Economics,* 20(1): 77-91.

[13] National Academies of Science (2004), *A Patent System for the 21st Century.*

[14] Priest, G. and B. Klein (1984), "The Selection of Disputes for Litigation," *Journal of Legal Studies,* 13: 1-56.

[15] Scotchmer, S., (2004), *Innovation and Incentives,* Cambridge, MA: MIT Press.

[16] Shavell, S. (1993), "The Optimal Structure of Law Enforcement," *Journal of Law and Economics,* 36: 255-87.

[17] Spier, K. (2005), "Litigation," *Handbook of Law and Economics,* M. Polinsky and S. Shavell ed., Elsevier, forthcoming.

[18] Waldfogel, J. (1998), "Reconciling Asymmetric Information and Divergent Expectations Theories of Litigation," *Journal of Law and Economics,* 41 (2): 451-76.

Back to Software "Profitable Piracy": The role of information diffusion and delayed adoption

Eric DARMON* Alexandra RUFINI† Dominique TORRE†

March 2008

Abstract

Can software piracy be profitable for a software editor? We tackle this issue in a simple model where software is an experience good and where the potential users of a software can choose to adopt or pirate a software or to delay their adoption. In that context, we show that a moderate piracy can be profitable for a software editor to foster users' adoption.

JEL Classification: D23, D42, L86
Keywords: piracy, experience goods, heterogeneous users, delayed adoption, signalling.

1 Introduction

The strategy of producers of digital goods (*e.g.* music, software) towards peer-to-peer communities has raised a huge theoretical debate about the existence of a so-called "profitable piracy". Because of the network externalities generated by digital products, and because of their specific cost function (high fixed cost and negligible marginal cost), a firm may find it profitable to distribute its product for free in order (notably) to increase buyers' valuation and to increase the price charged on regular consumers. For generic digital products, Gayer and Shy (2003) precisely analyze how peer-to-peer communities can be profitably used to enhance sales. In their model, potential adopters can either download or buy a digital good, these two goods are vertically differenciated and the downloaded product have a positive influence on the bought one (and conversely). Peitz and Waelbroeck (2006b) synthesize the literature in the case of digital products[1]. In the specific case of software, it has been first shown (Conner and Rummelt, 1991) that in the presence of network externalities, software editors may tolerate a "moderate" rate of piracy of their products. Shy and Thisse (1999) extend this result to a duopoly setting

*University of Rennes 1 - CREM - CNRS, 7 place Hoche, 35065 Rennes Cedex, France. E-mail: eric.darmon@univ-rennes1.fr

†University of Nice Sophia-Antipolis - GREDEG(DEMOS) - CNRS, 250 rue Albert Einstein, 06560 Valbonne, France. E-mail: alexandra.rufini@gredeg.cnrs.fr, dominique.torre@gredeg.cnrs.fr

[1]See Varian (2005) for a general economic approach of copyright and piracy; see also Qiu (2006) for a general equilibrium analysis linking software production and copyright protection.

and show how firms' incentives to tolerate piracy depend on the magnitude of the software network externality. This result has been further challenged by King and Lampe (2003) that pointed out that such trade-off could be not relevant in many cases. Yet, most of this literature emphasizes on one-shot games, where all potential users of a software decide simultaneously and once for all, whether to adopt or not. Doing so, such perspective neglects one key specificity of software goods: software are essentially *experience goods* that users need to sample before they know the exact utility they can derive from them. In that respect, the diffusion of software has to be analysed as a sequential process. Externalities between agents are here strictly informational and come from the application of the "sample effect" – first identified by Peitz and Waelbroeck (2006a) – to the particular context of software market[2].

We propose to come back to this debate about profitable piracy by stressing the particular relationship between information disclosure (about software quality) and piracy. Since users cannot perfectly assess the intrinsic quality of a software *ex ante* (experience goods), we consider a simple two-stage adoption setting where some early adopters can partially inform late adopters about the quality of the software, and where the firm can monitor various degrees of piracy of its products by implementing Digital Rights Management (DRM) tools. Within this setting, we show that a firm should accommodate a "moderate" piracy of its software to signal the quality of its software and hence increase its profits. Section 2 presents the model and establishes this result. Section 3 concludes.

2 The model

The firm. There is a single firm that produces software which provides both basic and advanced functionalities. The firm sells it at a price p_t at time $t, (t = 1, 2)$. To provide basic functionalities, the firms incurs a fixed cost C. The quality of advanced functionalities (denoted f) is a control variable for the firm. To keep things simple, we suppose that the quality level could take two discrete values $f = 0$ (low quality, *i.e.* no advanced functionalities) and $f = \bar{f} > 0$ (high quality) which imply an additional fixed cost $\chi(0) = 0$ and $\chi\left(\bar{f}\right) > 0$ respectively. The firm can implement various DRM technical solutions to monitor the piracy of its product. This choice is captured by λ_t, $(\lambda_t \in [0, 1])$: "hackers" of the software have a probability λ_t of being detected. We suppose that the level of monitoring is a control variable and that monitoring incurs no specific cost[3]. The objective of the firm is then to choose $\{p_1, p_2, \lambda_1, \lambda_2, f\}$ so as to maximize its intertemporal non-actualised profit defined by:

$$\begin{aligned} \pi(p_1, p_2, \lambda_1, \lambda_2, f) = \; & m_1^b(p_1, p_2, \lambda_1, \lambda_2, f)p_1 \\ & + m_2^b(p_1, p_2, \lambda_1, \lambda_2, f)p_2 - C - \chi(f) \end{aligned} \tag{1}$$

where $m_t^b(\cdot)$ is the quantity of software sold at time t.

[2]One exception is Chepalla and Shivendu (2005) that considers the effect of sampling (through piracy). Yet in their model, sampling is strictly personal (individual trial and error process) and there is no communication between agents about the quality of the software.

[3]This is a working assumption: if we find that a moderate piracy may be profitable for the firm when monitoring is costless, the same conclusion would hold more intensively when monitoring is costly. We have then chosen to skip the cost of monitoring so that our results are robust towards this cost.

Potential users. There exist m potential users of the software that we suppose uniformly distributed on the segment $[c, \bar{c}]$ where c_i figures the potential user i's cost of piracy. This piracy cost is incurred at the time the software is pirated. Users' heterogeneity relates both to technical factors (different abilities to pirate a software) or to psychological factors (different degrees of risk aversion for being detected, ethical/moral factors). These potential users have the opportunity to adopt the new software at time 1 or at time 2, either by buying it or by pirating it. This population is called early adopters When they do not adopt, they derive a utility $\bar{b}$ from the use of an old generation software (Reservation strategy). When they adopt the new software, they draw each period an instantaneous utility b from using the basic functionalities of the software and f from using its advanced functionalities. Basic functionalities are perfectly observable while advanced functionalities are initially not (since software are typically "experience goods"). At time 1, the instantaneous utility of these advanced functionalities is evaluated by potential users at its expected value $E(f)$,$\left(E(f) \leq \bar{f}\right)$. At time 2, the quality f is perfectly observed by agents who adopted the software at time 1. Information about quality is diffused to the other agents and improves their estimation of the quality of the software. This diffusion occurs through a word-of-mouth process, the efficiency of which depends on the number of early adopters (denoted m_1^a): it generates an externality from the early adopters to the remaining agents that we suppose linear.

At time 1, the objective of any potential user i is then to choose his present and future strategies in order to maximise the expected intertemporal utility U_i^1. According to the values of the parameters $\bar{b}, b, E(f), c, \bar{c}$ and to the observed control variables of the firm $(p_1, p_2, \lambda_1, \lambda_2)$, these possible actions and related payoffs are summarized in Table I.

Table I: User i's strategies at time 1

User's strategy	Time 1 action	Time 2 action as planned at time 1	Expected intertemporal (non-actualized) utility at time 1
1.1	Buying	-	$2(b + E(f)) - p_1$
1.2	Piracy	-	$2(b + E(f)) - \lambda_1 c_i$
1.3	Reservation	Buying	$\bar{b} + b + E(f) - p_2$
1.4	Reservation	Piracy	$\bar{b} + b + E(f) - \lambda_2 c_i$
1.5	Reservation	Reservation	$2\bar{b}$

To make the choices non trivial, we suppose *i)* $b \leq \bar{b}$, *i.e.* that the old software integrates both basic and advanced functionalities and *ii)* $\bar{b} \leq b + E(f)$, *i.e.* that there exists a non-negative price of the software such that all potential users do not choose once for all the Reservation strategy (user's strategy 1.5).

Since agents are not differentiated by their willingness to pay, the coexistence of buyers and non adopters at each time is excluded. Without any additional restriction, the set of the firm control variables then reduces from $(p_1, p_2, \lambda_1, \lambda_2, f)$ to $(p, \lambda_1, \lambda_2, f)$. Hence, at

time 1, since users' expectations about the current and future quality of the software are the same, they always prefer to buy the software immediately than later (thus eliminating strategy 1.3). Moreover, since we have supposed no direct network externality, the firm will have no interest to allow piracy at time 2 because the additional "hackers" at time 2 cannot have any positive influence on profit. Then, the interest of the firm is to announce at time 1 that the level of monitoring at time 2 will be sufficiently high to avoid incentives to pirate at time 2 (ensuring that user's strategy 1.4 is always dominated by strategy 1.2)[4]. Consequently, given their respective costs of piracy, potential users then select one of the three remaining strategies (1.1, 1.2 or 1.5).

Finally, there exist two possible distributions of users at time 1: total initial adoption (users are distributed among user's strategies 1.1 and/or 1.2) or partial initial adoption (users are distributed among user's strategies 1.2 and 1.5). These two distributions are the rational answers of potential users to the two strategies of the firm that we call Strategy A and Strategy B respectively. With Strategy A, the firm chooses its control variables $\{p, \lambda_1, \lambda_2, f\}$ such that high cost potential users choose to buy the software at time 1 while low cost ones choose to pirate it. With Strategy B, the firm determines its control variables such that high cost users choose Reservation at times 1 and 2 (user's strategy 1.5) whereas low cost agents choose piracy at time 1 (user's strategy 1.2). Here, the objective of the firm is to incite the agents who have chosen to reserve definitively at time 1 (user's strategy 1.5) to change their opinion with the new available information diffused by the early adopters and decide to buy the software at time 2. At time 2, the decisions of those agents who chose Reservation at time 1 are then detailed in Table II.

Table II: User i's strategies at time 2
when Reservation has been previously chosen at time 1

User's strategy	Time 1 action	Time 2 action	Expected intertemporal (non-actualized) utility at time 2
2.1	Reservation	Buying	$b + E(f) + (f - E(f))k(m_1^a/m) - p$
2.2	Reservation	Piracy	$b+E(f)+(f-E(f))k(m_1^a/m)-\lambda_2 c_i$
2.3	Reservation	Reservation	$\bar{b}$

Note that utilities are modified by the revision of the quality expectation. The magnitude of information diffusion depends i) on the proportion of early adopters and ii) on a constant k that measures the efficiency of the information diffusion process such that $0 \leq k \leq 1$.

There exist two variants of Strategy B. With Strategy B1, the optimal values selected by the firm for the control variables are such that all agents who chose Reservation at time 1 choose Buying at time 2. Yet, with Strategy B2, they split between buyers and "hackers". As they always correspond to non profit-maximizing outcomes for the firm,

[4]This remark explains also why the control variables p, λ_1 and λ_2 of the firm are known by potential users before they take any decision and why the announce of these control variables is credible.

other cases are excluded.

The structure of the game. The actions of the firm and the potential users can be depicted by a Stackelberg equilibrium where the firm plays leader:

- At time 0, the firm determines the level of $\{p, \lambda_1, \lambda_2, f\}$ maximizing the profit described by (1).
- At time 1, potential users formulate their intertemporal choices conditional to the information available on the control variables of the firm p, λ_1, λ_2 and implement their decisions related to time 1.
- At time 2, non adopters revise their time 1 decisions, according to the diffusion of information on the software quality at time 2 and implement these decisions concerning time 2.

The model is solved by backward induction. We are then able to prove the following proposition:

Proposition. *Through the implementation of fine-tuned protection devices, a strategy based on partial piracy may be profit-enhancing for the firm.*

Proof. see Appendix.

This proposition captures the following stylized fact: in some cases, it may be interesting for the firm to launch a new software with an initial adoption period where the rate of monitoring is not maximal. During this period, the firm tolerates piracy strategically. Such stage helps the firm maximize the disclosure of verifiable information about the quality of its software at the early stage of the diffusion process and helps selling it during the later stages of the diffusion process. It should be noted that piracy is beneficial here only if there exists a limited number of "hackers": this number must be sufficient to diffuse as broadly as necessary information on quality of the new software, while not excessive. This is why the firm needs to use fine-tuned protection devices (allowed by DRM implementations) so as to precisely control the piracy of its product.

Yet, this situation occurs when some conditions are filled ($\bar{c} < 4(b - \bar{b} + E(f))$ and conditions [I] to [III] in the Appendix). In particular, these conditions reveal that low quality basic functionnalities (b) or/and pessimistic expectations ($E(f)$) need to be challenged by the information revealed by hackers. At the same time, piracy costs ($\bar{c}$) should not be too low: if so, the information revealed by early adopters would encourage all those agents who did not adopt first, to pirate instead to buy further.

3 Concluding remarks

In this paper, we have analyzed how a software editor can strategically use piracy to increase the diffusion of its product and to maximize profit. In a simple setting where potential users can learn about the quality of a software (experience good) and where the firm can implement gradual DRM strategies (differing by the attitude of the firm

towards piracy), we have shown that, in some cases, piracy can even be profitable. Unlike some previous work, this conclusion is here only grounded on *informational* externalities between users, leaving voluntarily aside other types of externalities (compatibility effects, etc.). Besides, this conclusion is not dependant on particular assumption about the costs associated to the implementation of the piracy strategy: obviously, implementing DRM technical solutions may imply different production costs since creating a more and more "protected" software is more costly than producing a freely duplicable product. To avoid such exogenous dependence, we supposed that implementing all piracy policies imply the same cost. Future work should now consider software distribution strategy within an extended framework that could integrate beta-testing, together with product versioning.

Appendix

Sketch of the proof.

For each strategy A, B1 and B2:

- We determine the condition(s) on the distribution of users that make possible the adoption of strategies A, B1 or B2 from the producer's side.
- We express the profit and the optimal value of the control variables $\{p^*, \lambda_1^*, \lambda_2^*, f^*\}$. For the strategy B1, we define the condition on parameters such $0 < \lambda_1^* < 1$.
- Substituting the optimal values of the control variables in the condition(s) on the distribution of users, we determine the set of parameters for which each strategy is optimally used. We verify that the conditions on the distribution of users for strategy B1 are such that λ_1^* always takes an interior value.

Concluding the proof, we compare the strategies and find that there exist a non-empty range of parameters such that producers can rationally adopt B1 with $0 < \lambda_1^* < 1$.

Firm strategy A (all users adopt at time 1).

- With strategy A, the firm chooses driving potential users toward total adoption at time 1: potential users are then distributed among strategies 1.1 and 1.2, with

$$p \leq 2\left(b + E(f) - \bar{b}\right) \tag{2}$$

The distribution of users at time 1 is then depicted by Figure 1.

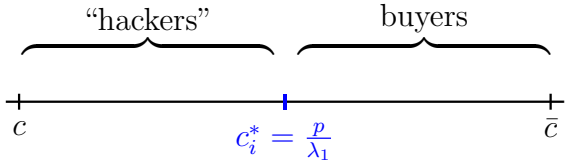


Figure 1: Distribution of users with strategy A

- According to (1), the profit is: $\pi^A(p, \lambda_1, \lambda_2, f) = [(\bar{c} - (p/\lambda_1))/(\bar{c} - \underline{c})]mp - C - \chi(\bar{f})$. From the FOC on the profit and given the interval definition of λ_1, we deduce that $\lambda_1^* = 1$ and $p^* = \bar{c}/2$. λ_2 has to be chosen such that potential users do not pirate at time 2, *i.e.* $1 - (b + E(f) - \bar{b})/\bar{c} \leq \lambda_2^* \leq 1$. Since agents only buy at time 1 without knowing the exact quality of the functionalities, the value $f^* = 0$ maximizes the profit π^A.
- Given the optimal value of p^*, condition (2) finally becomes:

$$\bar{c} \leq 4(b - \bar{b} + E(f))$$

Firm strategy B1 (partial adoption at time 1, only buyers at time 2).

- With strategy B1, the firm compels potential users to split between strategies 1.2 and 1.5 / 2.1. The values of the control variables of the firm $p, \lambda_1, \lambda_2, f$ are then such that:

$$p > 2\left(b + E(f) - \bar{b}\right) \tag{3}$$

$$c < \frac{2\left(b + E\left(f\right) - \bar{b}\right)}{\lambda_1} < \bar{c} \tag{4}$$

$$p \leq \frac{2(b + E\left(f\right) - \bar{b})(f - E\left(f\right)) + \lambda_1((\bar{c} - c)(b + E\left(f\right) - \bar{b}) + (f - E\left(f\right))kc)}{\lambda_1(\bar{c} - c)} \tag{5}$$

$$p \leq \frac{2\lambda_2(b + E\left(f\right) - \bar{b})}{\lambda_1} \tag{6}$$

The distribution of users at time 2 is then depicted by Figure 2.

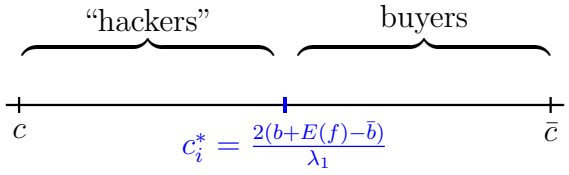


Figure 2: Distribution of users with strategy B1

- In this case, whatever the optimal price p^*, the firm has to persuade agents who have not adopted at time 1 to buy the software at time 2. For that, the software has to be endowed of high quality functionalities, *i.e.* $f^* = \bar{f}$. Formally, the profit is: $\pi^{B1}(p, \lambda_1, \lambda_2, f) = [(\bar{c} - ((2(b + E(f) - \bar{b}))/\lambda_1))/(\bar{c} - c)]mp - C - \chi(\bar{f})$. From the FOC, we deduce that: $p^* = \left[b + E(f) - \bar{b} + k\left(f - E(f)\right)\right]/2$, $f^* = \bar{f}$, $\lambda_2^* = 1$ and $\lambda_1^* = 4(b + E\left(f\right) - \bar{b})(f - E\left(f\right))k/[(\bar{c} + c)(f - E\left(f\right))k - (\bar{c} - c)(b + E\left(f\right) - \bar{b})]$. Since all parameters are positive, λ_1^* is positive also. Then the corner solution $\lambda_1^* = 1$ is excluded if $4(b + E\left(f\right) - \bar{b})(f - E\left(f\right))k/[(\bar{c} + c)(f - E\left(f\right))k - (\bar{c} - c)(b + E\left(f\right) - \bar{b})] < 1$.

- Hence, when $(p, \lambda_1, \lambda_2, f)^*$ are put into (3), (4), (5) and (6) the following conditions emerge:

$$k \geq \frac{3(b + E(f) - \bar{b})}{f - E(f)} \textbf{ [condition I]}$$

$$\bar{c} > \frac{(c - k(\bar{f} - E\left(f\right)))(b + E\left(f\right) - \bar{b} + k(\bar{f} - E\left(f\right)))}{(b + E\left(f\right) - \bar{b} - k(f - E\left(f\right)))} \textbf{ [condition II]}$$

One can verify that the combination of **[condition I]**, **[condition II]** finally makes always true the condition ensuring that λ_1^* takes an interior value ($4(b + E\left(f\right) - \bar{b})(f - E\left(f\right))k/[(\bar{c} + c)(f - E\left(f\right))k - (\bar{c} - c)(b + E\left(f\right) - \bar{b})] < 1$). Finally, since providing high quality functionalities is costly, the strategy B1 ensures a non negative

profit and is then chosen by the firm if:

$$\chi\left(\bar{f}\right) \leq \frac{m(b + E\left(f\right) - \bar{b} + k(\bar{f} - E\left(f\right)))^2}{4k(\bar{f} - E\left(f\right))} - C \text{ [condition III]}$$

Firm strategy B2 (partial adoption at time 1, buyers and "hackers" at time 2).

- When the strategy B2 is used, the firm chooses to distribute agents among user's strategies 1.2 and 1.5 / 2.1 and 2.2. This case occurs if conditions (3) and (4) still hold, as in Strategy B1 and if:

$$\frac{2\left(b + E\left(f\right) - \bar{b}\right)}{\lambda_1} < \frac{p}{\lambda_2} \leq \bar{c} \tag{7}$$

The distribution of users at time 2 is depicted by Figure 3.

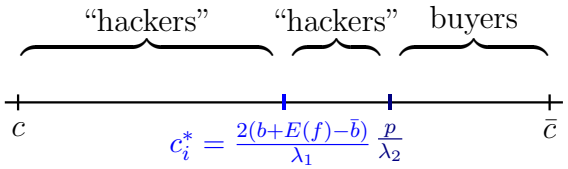


Figure 3: Distribution of users with strategy B2

- Formally, the profit is: $\pi^{B2}(p, \lambda_1, \lambda_2, f) = [(\bar{c} - (p/\lambda_2))/(\bar{c} - c)]mp - C - \chi(\bar{f})$. Whatever λ_1 (since λ_1 does not appear in the equation profit), the firm chooses the optimal values of $(p, \lambda_2, f)^*$ from the FOC, *i.e.* $p^* = \bar{c}/2$, $f^* = \bar{f}$ and $\lambda_2^* = 1$. λ_1^* has not an unique possible value but has to be chosen in the interval $[4(b + E\left(f\right) - \bar{b})/\bar{c}\ ,\ 4(b + E\left(f\right) - \bar{b})(\bar{f} - E\left(f\right))k/((\bar{c} - c)(\bar{c} - 2b - 2E\left(f\right) + 2\bar{b}) + 2(\bar{f} - E\left(f\right))kc)]$ to fulfill (4), (5) and (7).

- Given the optimal value of p^*, (3) becomes $\bar{c} > 4\left[b - \bar{b} + E(f)\right]$ while (4), (5) and (7) always hold since the optimal value of λ_1^* fulfills these conditions.

  Similarly to [**condition III**], the strategy B2 is activated by the firm if:

$$\chi(\bar{f}) \leq \frac{m\bar{c}^2}{4(\bar{c} - c)} - C \text{ [condition IV]}$$

Conclusion of the proof.

If $\bar{c} \leq 4\left[b - \bar{b} + E(f)\right]$, two strategies are available: strategy A and strategy B1. The comparison of profits shows that strategy A is always prefered by the producer to strategy B1 even if there is no costs incurred by the advanced functionalities.

If $\bar{c} > 4\left[b - \bar{b} + E(f)\right]$, the firm chooses strategy B1 and then allow some piracy (with this strategy λ_1 takes an interior value) if [**condition I**], [**condition II**] and [**condition III**] hold. If [**condition II**] does not hold, the optimal strategy is strategy B2 but only if [**condition IV**] holds; otherwise the firm will not produce the software.

References

Chellappa, R.K., and S. Shivendu (2005) "Managing Piracy: Pricing and Sampling, Strategies for Digital Experience Goods in Vertically Segmented Markets" *Information Systems Research* **16**, 400-417.

Conner, K.R., and R.P. Rummelt (1991) "Software piracy: an analysis of protection strategies" *Management Science* **37**, 125-139.

Gayer, A., and O. Shy (2003) "Internet and peer-to-peer distributions in markets for digital products" *Economic Letters* **81**, 197-203.

King, S.P., and R. Lampe (2003) "Network Externalities, Price Discrimination and Profitable Piracy" *Information Economics and Policy* **15**, 271-290.

Peitz, M., and P. Waelbroeck (2006a) "Why the music industry may gain from free downloading - The role of sampling" *International Journal of Industrial Organization* **24**, 907-913.

Peitz, M., and P. Waelbroeck (2006b) "Piracy of digital products: A critical review of the theoretical literature" *Information Economics and Policy* **18**, 449-476.

Qiu, L.D. (2006) "A general equilibrium analysis of software development: Copyright protection and contract enforcement" *European Economic Review* **50**, 1661-1682.

Shy, O., and J-F. Thisse (1999) "A Strategic Approach to Software Protection" *Journal of Economics and Management Strategy* **8**, 163-190.

Varian, H. (2005) "Copying and Copyright" *Journal of Economic Perspectives* **19**, 121-138.

Dépôt légal : 2008 – 4ème trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 1242-5125 ENST S (Paris) (France 1983-9999)