Three Essays on Public Economics and Strategic Behavior

PhD Thesis in Economics

António Freitas

Thesis Supervisor: Prof. Inés Macho Stadler

International Doctorate in Economic Analysis (IDEA)

Universitat Autònoma de Barcelona

Barcelona, Spain
Department of Economics and Economic History

May, 2012.

DEDICATION

To my family, who always guides me to become a better person.

A toda a minha familia, que sempre me guia para ser uma pessoa melhor.

ACKNOWLEDGMENTS

This space is dedicated to those who contributed, professionally and personally, to this work. All of them I leave here my sincere thanks for the support in times when I needed it so much.

Firstly, I thank my supervisor Inés Macho Stadler for her outstanding support and guidance in this long path. During the ellaboration of this thesis, Inés was very involved in my research work. Inés was always available to meet me, as well as to listen to my regrets in times of less optimistism or doubt. Thank you Inés, you have become a role model to me. You have shown me so many important values such as commitment, hard work, optimism and, last but not least, friendship. I will keep them with me always.

In addition, I am thankful to Professors David Pérez Castrillo, Pau Olivella, Pedro Rey Biel, Stefano Trento and Xavier Martinez Giralt, who provided significant contributions, suggestions and important advices. A special mention to Professors Joan Calzada, Jose Sempere, and Albert Banal Estañol for accepting the invitation to be part of the jury.

I am also very grateful to all the faculty members and teachers of the Economics Department, who contributed to the acomplishment of this thesis. Among others, I would particularly thank my co-authors Inés and Michele. Also I could not do this thesis without the help of our marvellous administrative team of the Economics Department, Mercé and Àngels, who were tireless in helping with any problem one could have, always with a smile.

My colleagues have been crucial in doing this journey, they made it seem so much easier. In particular I thank Grisel Ayllon for being the best mom in Barcelona and my true friend, to Joaquin for our long talks, to Marion, Tatjana and Nuno, among others.

To all my family, I thank you for being always by my side. Firstly my parents, Pai e Mãe, thank you for the very frequent calls to know how I am and always giving me the spirit to go on everyday. Mom thank you for your love and passion in telling me to go on, and Dad thank you for always suggesting a book about one topic or another, I have learned from you the need to go beyond my knowledge limits. To my brother and sister Nuno and Alexandra, thank you for being a role model to your youngest brother and for giving me four worderful nephews.

I would also like to thank my friends spread between Lisbon, Barcelona and Mexico. I specially thank Fermín, for being my best friend and for always being there when I needed, even at the distance. To my second mom Lá-Salette, my "godmother" Gabriela, David, Juan, André, Carlos, Ricard, Montse, Toni, Darwin, Fred, and many other friends. Thank you for believing in me, you are the non-blood family I found.

Last but not least, I thank Fundação Ciência e Tecnologia, from the Portuguese Science Ministry, for trusting my skills and providing financial support throughout 4 years with the Grant SFRH/BD/40182/2007. Without it I could not have done this work.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1.

INTRODUCTION

Public Economics consists in identifying and evaluating the effects of public intervention on economic outcomes, where economic efficiency and equity stand out as important inbedded concepts. The main purpose of public policy coincides with the aim of this thesis, that is, to identify situations where resources are not being used efficiently, to assess the role of a government or regulator and to evaluate to which extent authorities can intervene to minimize the effects of market failures, to maximize social welfare and to achieve dynamic efficiency. Also, when examining the role of public policy, one must account for the strategic behavior of economic agents, specifically when individuals and firms are rational and decide following their self-interest or even in cases when the rationality of individuals is bounded.

In my thesis, the central questions apply to the general topic of innovation, under an industrial organization approach. I show how different regulatory regimes or policies can shape the incentives of agents or firms to invest more or to invest differently, in either dedicating time to train scientists, investing in intellectual property rights, or investing in online network capacity. All these decisions shape the innovation path of these three different economic sectors and have an impact on the welfare of consumers.

In Chapter 2 of this dissertation, I consider a framework where universities and research institutions contribute to the quality of future researchers and scientific projects. Senior, experienced scientists must allocate time between doing research and training junior scientists where a system of apprenticeship is very important in passing

knowledge and acquiring expertise. I evaluate the effect of this time allocation decision on the quality of juniors and on the value of science projects.

More precisely, I consider a model a senior scientist who must choose how much time to spend performing direct research and how much time to spend training to a junior scientist, who also participates in the project. Both senior and junior do not have information about the junior's innate ability. On one hand, working by herself can be extremely rewarding to the project. On the other hand, training the junior adds value to the project as well, since he can increase his productivity and make better contributions to the project, increasing its value. The more able is the junior, the higher is his scientific contribution to the project.

I show that the time allocation decision of the senior depends on the characteristics of the project, on her concern for training and the expected innate ability of the junior. Also, when players have no information about the junior's innate ability, there is a mismatch: the senior may provide excessive training to the least able scientists and insufficient training to the most talented juniors. Additionally, I show that there are cases where the senior scientist is willing to spend time performing a selection process to filter more talented junior scientists to integrate in the project.

I analyze the role that regulation can play in defining both the value of scientific projects and the population of future independent scientists. The implementation of training programs in earlier education and tougher selection processes to attract high-ability junior scientists for research under supervision, as well as attractive training conditions, can be effective measures to attain it.

In Chapter 3, I investigate the implications on patent strategy of pharmaceuticals

when brand pharmaceutical firms are allowed to settle patent disputes with generic firms through pay-for-delay settlements. Pay-for delay settlements are agreements where the incumbent pays the entrant to stay out of the market.

In the United States, the Hatch-Waxman Act in 1984 made possible that generic firms enter brand drug markets before the patent of the brand drug expires. Since the patent is still due to expire, incumbent firms can and often react to generic entry and enter into patent dispute. In recent years, several patent disputes have been solved out of court, before reaching a trial, in the form of *pay-for-delay* settlements, that is, by delaying the entry decision of generic firms in exchange of a compensation. Despite antitrust allegations on behalf of competition authoriries, U.S. federal courts have had a lenient approach and have not ruled them as illegal. In the U.S. there is a large debate on whether pay-for-delay settlements should be allowed or not, a debate in which the Federal Trade Comission actively supports a complete ban on this type of settlements.

In this context, I analyze whether pay-for-delay settlements are having an effect on what brand drug firms are patenting. In my model of strategic behavior endogenous dispute, I compare the patenting decision of pharmaceutical firm in two different legal frameworks, one when pay-for-delay settlements are allowed and one where they are not. More specifically, I study the patenting decision of one incumbent, brand drug firm, that operates in a market protected by one patent and faces potential entry. In case of entry, the brand drug firm may accommodate entry or may defend the patent in a legal dispute with the generic firm. The decision of the brand drug firm is whether to develop a new patent on a new drug or invest in improving protection of the existing drug.

My analysis shows that, when pay-for-delay settlements are allowed, generic firms enter the existing drug market more frequently and brand drug firms tend to direct the patent decision towards protection of existing drugs, in detriment of new drug development. This occurs when the drug market is attractive enough so that a generic enters under pay-for-delay settlements but does not when pay-for-delay settlements are not allowed, and when trial costs are higher than settlement costs. Hence, pay-for-delay settlements can shift pharmaceuticals' strategies away from new drug discovery.

Finally, in Chapter 4 of this dissertation I study the market for internet services and, more precisely, the investment incentives of internet service providers when they are allowed the possibility to discriminate, that is, to offer a tiered online service to content providers.

In my model, the internet market is composed of three types of players, internet service providers, content providers, and users. Internet service providers are networks that work as platforms to connect users that wish to access contents online. The users' decision of which content to acess depends on the preference for the content but also on the waiting time to acess the content.

While network neutrality is based on the principle that all traffic in internet should be treated in an identical way, network discrimination implies internet providers can manage traffic and offer a tiered internet connection service to producers of content. This means that content providers that pay a fee obtain a better quality delivery service, a priority service. The privileged position of internet providers as managers of data traffic has triggered a strong debate on network neutrality, both in the United States and in Europe.

While in the U.S. the discussion is still under way to impose neutrality in online service or not, Europe's position has been that internet service provider competition should mitigate abbusive behavior of service providers and be a driver of investment in capacity, regardless of the network regime. Building upon the work of Choi and Kim (2010), who takes into account a monopolist internet provider, no theoretical framework has introduced competition between internet providers joint with traffic management and network congestion.

The analysis shows when the incentives of internet service providers to invest in network capacity increase in one regime with respect to the other. I show that when service providers can prioritize and always have the incentive to do so, users tend to switch from the large provider to the small one. Additionally, if internet providers charge a sufficiently high fee to content providers for the priority service, they always have less incentives to increase network capacity when compared to the network neutrality regime. This implies that competition does not necessarily guarantee higher investment in network capacity.

CHAPTER 2.

ON THE JOINT PRODUCTION OF RESEARCH AND TRAINING

2.1 Introduction

It is widely accepted that universities and research institutions have the responsibility to produce science. However, there is another task of great importance to our society's advancement of knowledge: training the new generations of researchers. In this paper, we consider senior scientists to be involved both in doing research and in providing training to junior scientists, as in a system of apprenticeship. Understanding the allocation of time among the two activities is of great interest because the training of junior researchers needs to be performed by the people who know how to do research, and this is crucial in assuring a high-quality research workforce for the future. However, there are voices that point out that our research institutions may be failing in this dimension, meaning that there is a shortage of time devoted to training scientists able to perform outstanding independent research in the future.[1] In this paper, we propose a model to address this problem, to discuss the allocation of time between research and training the next generation of researchers and to discuss the problems that may arise.

Our motivation comes from two facts. On the one hand, it is a documented fact that the most prominent candidates who attain an independent research status are PhDs and postdoctoral researchers, who thrive through more experience and skills either in academics or in industry (Cech and Bond, 2004). The literature on higher

[1] Obviously, an alternative to training one's own researchers is to attract researchers trained elsewere. While this is an interesting idea, we choose to ignore this topic in this paper.

education, human resource management and mentoring extensively recognizes the effects of training by senior staff in the competence, productivity, career development and independent skills of young professionals, both in industry and academia. The *student-supervisor* relationship is the most critical issue affecting the quality of the PhD training (which affects both eventual job placement and success in obtaining a degree). In this process, it is natural that doctoral students hold expectations with respect to the role of their supervisors. The most important expectations are guidance in the early days of obtaining a PhD, knowledge about the area they are working in, and most importantly involvement with their work (Pole *et al.*, 1997). On a postdoctoral level, Vogel (1999) reports the experience of a principal investigator (PI) supervising postdocs in an internationally appraised lab.[2] She states that the PI's key to producing successful and high-quality junior scientists is to provide them with original ideas and orientation, to encourage strong participation in the projects, and to listen to them to assess their skills, motivations and ambitions.

On the other hand, training problems persist on a global scale. Student doctoral attrition remains a common problem in PhD training, and this is estimated at approximately 50% on the U.S. (Lovitts, 2001). In a case study about former students who spent at least two years in a PhD program, Golde (2000) identifies a lack of support and guidance from supervisors as one of the causes of attrition. The author is also able to identify characteristics of a good supervision: the amount of time spent, the quality of interactions between student and supervisor, and an interest in the student's work

[2] A PI is a head researcher and author who supervises doctoral students, conceives ideas and conducts projects that may include collaborating with research assistants (Armbruster, 2008).

are important to guarantee training success. Accessibility seems to be an important issue as well.[3] Training problems also occur at the postdoctoral level. Puljak (2006) reports that the most common complaint in postdoctoral training is ironically, a the lack of postdoctoral training. Postdocs join a research lab and, shortly thereafter, many realize that they are on their own. It has also been identified that some advisors tend to take over the design of experiments, making postdocs feel like they are overeducated technicians.[4]

We study this issue by building a multitask model that examines the incentives of a senior scientist to provide training to a junior scientist. The senior scientist chooses the time to allocate to her own research and the time to train to the junior scientist under her supervision. We then evaluate the impact of time allocation on the level of research and the final skills of the junior scientists as a researcher. The junior scientist is not an active player in our model (he does not make any decisions), but has a productive role in the project (he contributes to its final value). In addition, we assume that the junior scientist's final capabilities are not only affected by the training received from the senior scientist but are also affected by his innate ability. On this respect, we abstract from information features: senior and junior scientists have the same information about the junior scientist innate ability.[5]

[3]It seems that there can be a mismatch in the perceptions of the supervisor and of doctoral students with respect to accessibility. In a study on the provisions of PhD training in biomedical research PhD programs, virtually all supervisors reported meeting frequently with their students, whereas 1/4 of the students reported problems in accessing their supervisor (Frame and Allen, 2002).

[4]Nerad and Cerny (1999) also survey the perspective on postdoctoral employment in the U.S. and report that there exists a generalized discontent on behalf of postdoctoral researchers. The length of postdoctoral appointments has increased and these appointments are increasingly being seen as 'holding base', rather than being an important step in a young researcher's career.

[5]This does not mean that the junior scientist's innate ability is public information. It is ex-ante unknown by both participants. Even though a student is selected to participate in a graduate

Not surprisingly, when we analyze the senior scientist's allocation of time between research and training, we find that when she has more time available this results in more time allocated to both tasks. Also, we show that when there is an increase in the innate ability of the junior scientist, an increase in the importance that training has on the senior scientist's utility function, or a decrease in the productivity of the senior scientist in the project then there is a tendency to increase the time allocated to training. We also discuss the final capability of the junior scientist and the final value of the scientific project in different scenarios. Most interestingly, we show that ignorance about the true innate ability of the junior scientist may lead to more training for less able junior scientists, while there is a tendency toward an under-investment in training for the most talented ones.

As a robustness check, we consider two extensions. First, we consider the case where the senior scientist can also spend time selecting a better junior scientist. Second we examine the case when the senior scientist chooses how much time she works, that is, the total amount of time spent on both tasks.

We also discuss possible policy instruments for a regulator who is concerned with maximizing the value of projects and attaining highly qualified scientists in a desired proportion. We highlight that the implementation of training programs in earlier education and tougher selection processes to attract high-ability junior scientists for research under supervision, as well as attractive training conditions, can be effective measures to attain it.

programme or in a lab according to a GRE score, and other internal admission criteria of a department, significant uncertainty remains in predicting if a student has the potential to become a successful independent researcher (Lovitts, 2005).

Following the work of Holmstrom and Milgrom (1991), where the authors propose a principal-agent model where the principal wants the agent to perform multiple tasks, several papers have considered the incentives for scientists to perform different tasks. For example, Lacetera and Zirulia (2008), in a context of corporate science with a great deal of competition, propose a model to explain the optimal choice of an effort to do applied research and an effort to do basic research. They analyze the strength of incentives in the effort allocation decision of the scientist and the effects of different levels of competition. In Banal-Estañol and Macho-Stadler's work (2010), the authors present a model of incentives of a researcher who can choose to either allocate time between undertaking a new research idea or developing an existing one that will deliver immediate commercial benefits. In the same branch of the literature, Walckiers (2008) argues about whether it is more attractive for a university to produce both research and teaching. The author conducts his analysis in a contractual setting between the university (principal) and the academic/scientist (agent) and studies the incentives for university scientists to perform either one of the tasks or both of them. In contrast to Walckiers (2008), where the agent does teaching at the undergraduate level, we consider training at the graduate level which implies that there are complementarities among the two tasks. Walckiers (2008) uses an adverse selection framework, where researchers differ on their preference for both tasks[6] and he shows that it can be optimal to produce research and teaching in the same institution (bundling the two tasks).

This paper is organized as follows. Section 2.2 describes the model and analyzes

[6]In our model, we could also discuss the researcher preferences, but this is not the main aspect of the analysis.

the equilibrium allocation of time to the two tasks: research and training. It also provides the comparative statics of the equilibrium efforts with respect to the parameters of the model. In Section 2.3, we evaluate and draw the patterns that the project's expected value and the junior scientist's final capability follow. We also present the ex-post ability of the junior scientist and the role of imperfect information in the distortions with respect to the full information and efficient outcomes. In Section 2.4 we perform a robustness check by considering two possibilities. First, we consider that the senior scientist can choose the total amount of time to exert in both tasks. Second, we consider the incentives for a senior scientist to spend time in previous activities that allow her to know more about the innate ability of the junior scientist. In Section 2.5 we discuss some policy instruments that may change the time allocation. In Section 2.6 we conclude. All proofs are remitted in the Appendix.

2.2 Basic Model

We consider a senior scientist who is in charge of a research project and allocates her time between research and the training of a junior scientist under her supervision. We denote the research effort by e_R and the training (guidance or education) effort by e_G and assume that the senior scientist has limited time t to allocate to these tasks. Formally, the senior scientist's time constraint is written as:

$$e_R + e_G = t.$$

In Section 2.4, we study how the available time t that the senior scientist works is determined. For now, we assume that $t > 0$ is exogenously given.

In our model, the junior scientist does not make any decisions. He is endowed with an innate ability $\hat{a}$, *ex-ante* unknown by all the players. We assume that there is a population of junior scientists with different innate abilities. The innate ability of the junior scientist who works with the senior takes a value in the interval $[\underline{a}, \bar{a}]$, with $\bar{a} > \underline{a}$, and the expected innate ability is $E(a)$.[7]

The senior scientist's vector of efforts affects two outcomes: the quality (the value) of the research project and the final capability of the junior she is training.

The final capability of the junior scientist depends on his innate ability and on the senior scientist's educational effort. Our view is that education provided by the senior is a necessary input to develop the junior's scientific capability. The junior scientist's final capability, denoted by q, is a function of his *true* innate ability $\hat{a} \in [\underline{a}, \bar{a}]$ and the training he receives e_G, and it is defined as follows:

$$q = \hat{a} e_G. \tag{2..1}$$

Without training, even the most gifted junior scientist will not be able to acquire the capability to work on the research project in a profitable way (and maybe run a research project in the future).

The project's value depends on the direct research effort exerted by the senior scientist and on the junior scientist's final capabilities. The scientific value of the senior's project is given by:

$$v = \alpha e_R + \delta q e_R \tag{2..2}$$

[7] In our model, there is always symmetric information about the junior scientist's innate ability. Under complete information senior and junior know that his innate ability is $\hat{a}$; under ignorance they expect it to be $E(a)$.

where α is the productivity of the senior's research time e_R, and δ captures the synergies of working together with a junior scientist of capability q. In this sense, senior and junior scientists provide complementary inputs to the research project. Researchers may have different projects defined by (α, δ) and we will discuss this further on. The level v may represent the publications obtained, patents achieved or other results of the discoveries. Note that by following this functional form, no value will be produced from the project if the senior scientist does not provide any research effort.

We assume that the senior scientist's utility function combines the project's value and the junior's final capability. Formally,

$$u(v, \beta, q) = v + \beta q.$$

The project's value is included in the senior's preferences because it is a verifiable outcome that determines the senior scientist's payoff. It is also a proxy for the usual argument of peer recognition and the "puzzle joy" (Stephan and Levin, 1992). The junior scientist's final capability enters the senior's utility in a proportion β, which represents the relative appreciation of the training outcome. We may also interpret this second term of the senior's utility as a concern with reputation associated to having a network and disciples who excel in the profession.[8]

Given the parameters $(\alpha, \delta, \beta, t)$ and the ex-ante expectation about the junior

[8] Our model aims to encompass the fact that both senior and junior scientists benefit from the training relationship. For the senior scientist, training increases visibility and reputation when the young professional is a productive member. Therefore, she earns more respect from the organization by developing the trainee (Kram, 1983). This model also includes the trainer's inner satisfaction in passing along knowledge (Levinson et al., 1978). For the junior scientist, the benefits include learning technical aspects of the profession, developing writing and critical skills, defining career perspectives, performing research collaborations (Kram, 1983) and receiving an important push toward building networks (Kram and Isabella, 1985).

scientist's innate ability a with $a \in \{\hat{a}, E(a)\}$, the senior scientist chooses the optimal allocation of time between e_R and e_G that maximizes the ex-ante (expected) value of her utility:[9]

$$\underset{e_R, e_G}{Max} \ \{\alpha e_R + \delta a e_G e_R + \beta a e_G\}$$

$$s.t. \qquad e_R + e_G = t$$

$$e_R \in [0, t]$$

$$e_G \in [0, t]$$

From the solution to this problem we obtain the result that follows.

Lemma 2..1 *Given $(\alpha, \delta, \beta, t)$ and the innate ability of the junior scientist a, with $a \in \{\hat{a}, E(a)\}$, the senior scientist's allocation of time among the tasks of research and training is:*

a) When $\beta > \delta t$ and $a > \frac{\alpha}{\beta - \delta t}$,

$$e_R^* = 0 \ \text{and} \ e_G^* = t$$

b) When $a < \frac{\alpha}{\beta + \delta t}$,

$$e_R^* = t \ \text{and} \ e_G^* = 0$$

c) Otherwise,

$$e_R^* = \frac{t}{2} + \frac{\alpha - \beta a}{2a\delta} \ \text{and} \ e_G^* = \frac{t}{2} - \frac{\alpha - \beta a}{2a\delta}$$

[9]Note that v is a linear function of a that allows us to use a simplification where we write the ex-ante value of the project as a function of a, $a \in \{\hat{a}, E(a)\}$. This allows us to consider at the same time the cases where the information about the junior scientist's innate ability is perfect or imperfect.

The senior scientist's allocation of time for the interior solution depicted in Lemma 2..1

$$\left(e_R^* = \frac{t}{2} + \frac{\alpha - \beta a}{2a\delta}, e_G^* = \frac{t}{2} - \frac{\alpha - \beta a}{2a\delta}\right)$$

depends on all the relevant parameters. It shows the deviation from the half-half distribution of time t as a function of the senior scientist's effectiveness in the research process (α), the complementarity among the senior's and junior scientist's participation (δ), the expected innate ability of the junior (a) and the senior's concern about the junior's training (β). Lemma 2..1 also shows that when the junior has a low expected ability he does not receive any training. It also shows that when the senior scientist's concern about the junior's training is high as compared to the time available and the complementarity ($\beta > \delta t$) then the possibility exists that she decides to train only. Note for example, that if $\delta = 0$ (and $a \neq \frac{\alpha}{\beta},$), i.e., there is no effect of the junior scientist toward the result of the research project, then time will only be allocated to training.

Corollary 2..1 *For the combination of parameters satisfying* $\alpha \in [a(\beta - \delta t), a(\beta + \delta t)]$ *(region c) in Lemma 2..1), the static comparative of the efforts is presented in Table 1:*

	α	β	δ	t	a
e_R^*	$+$	$-$	$-$ *iff* $\alpha > a\beta$	$+$	$-$
e_G^*	$-$	$+$	$+$ *iff* $\alpha > a\beta$	$+$	$+$

Table 1

As expected, the senior scientist's research effort (resp., training effort) increases (resp., decreases) when α increases, β decreases or a decreases. Both efforts go in

different directions when δ increases, and the direction of the change depends on the sign of $\alpha - a\beta$. In addition, both efforts increase with the time t the senior scientist has to work.

The effect of a in equilibrium is in accordance with stylized facts. In a 5 $\frac{1}{2}$ year longitudinal investigation with 233 PhD students (Paglis, Green, and Bauer, 2006, which was an extension of a similar study by Green and Bauer, 1995), the effects of supervisory mentoring of advisors on PhD students in the applied sciences were analyzed. The results show that supervisory mentoring increases the productivity and the self-efficacy of PhD students. Most importantly, a positive relationship between student potential (ability, experience and commitment to the training) and the extent to which training functions are provided by the faculty advisor is identified. That is, students who show more promise to be productive researchers receive more supervisory training and mentoring from the advisor.

If we consider, ceteris paribus, a constellation of projects defined by the pair (α, δ), for the projects whose success is heavily based on the time the senior scientist spends on it, training is weak.[10] Also, if the direct participation of the senior scientist in the project is very important (α) and the complementarity (δ) among junior and senior participation increases, then the allocation of time to training also increases.

With respect to β, when $\beta = \frac{\alpha}{a}$, the senior scientist will exert an equal effort toward both tasks, $\frac{t}{2}$. For a higher β, the senior is relatively more concerned about training and the effect on her reputation, so she exerts a greater amount of effort to

[10]Note that in our model any project can provide the same quality of training since we assume that training may be field-specific but not project-specific.

training in detriment to research. Because the efforts are linear in this parameter, we observe a constant rate of substitution. Note that even if $\beta = 0$ (in which case we will have $\delta t \geq \beta$), the senior scientist may be interested in allocating time to training. To make this point clear, let us take the example $\delta = 1$. Then the important comparison is in between α and a. If the productivity of the senior scientist is high, $\frac{\alpha}{t} > a$, she just allocates time to research and the project runs exclusively on her effort. If α is low when compared to a then she will allocate time to training because the complementarity of her research effort with the junior scientist will be the motor of the project.

In Figure 2..1 we represent the senior's effort equilibrium allocation to the two tasks in the space (a, t) by representing the iso-effort curves (the dotted lines). Keeping constant the remaining parameters, any combination (a, t) provides an optimal combination of efforts. A low amount of time tends to concentrate the senior scientist's attention on one of the tasks: if a is also low the task will be research, if a is high the task will be training. In the interior solution, the shape of the iso-effort curves shows the conflicting effects between t and a in the optimal allocation of time. As seen in Corollary 2..1, both parameters are aligned in their effect on training, but induce different behaviors in their effect on research, which translates in the different slopes of the iso-effort curves. For example, as a increases and t decreases, less time is devoted to research but training may decrease or increase. If a and t increase then training also increases but research may decrease or increase depending on the relative change of both parameters. Finally, note that as δ decreases, $\frac{\alpha-\beta a}{a\delta}$ and $\frac{\beta a-\alpha}{a\delta}$ increase and there are more corner solutions where the senior scientist allocates her time only to one of the tasks.

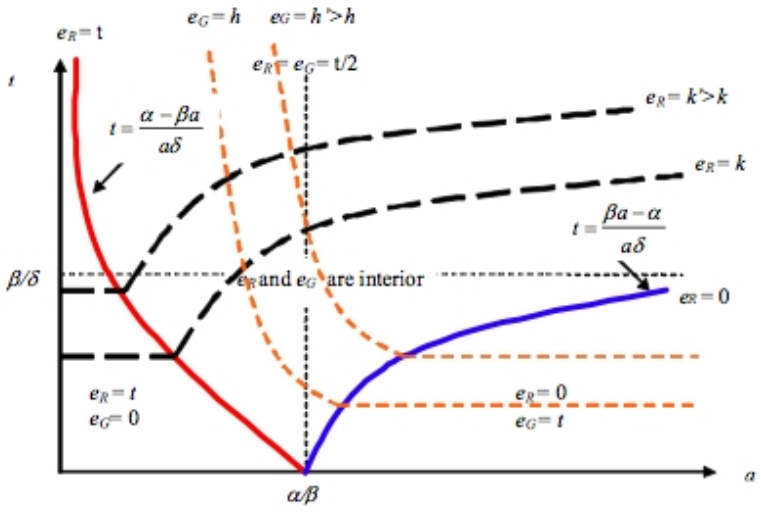

Figure 2..1: Optimal allocation of time

2.3 Expected Project Value and Junior Scientist Capability at the Optimal Time Allocation

Let us now focus on the effect that the optimal time allocation to both tasks has on the expected project value, v, as well as on the expected junior scientist capability, q.Using the results of Lemma 2..1, we compute the *ex-ante* (expected) equilibrium levels of v and q. We find that:

a) When $\beta > \delta t$ and $a > \frac{\alpha}{\beta - \delta t}$,

$$v^*(a) = 0 \text{ and } q^*(a) = at$$

b) When $a < \frac{\alpha}{\beta + \delta t}$,

$$v^*(a) = \alpha t \text{ and } q^*(a) = 0$$

c) Otherwise,

$$v^*(a) = \frac{(a\delta t + \alpha)^2 - (\beta a)^2}{4a\delta} \text{ and } q^*(a) = \frac{ta}{2} - \frac{\alpha - \beta a}{2\delta}$$

When the parameters satisfy $\alpha \in [a(\beta - \delta t), a(\beta + \delta t)]$ (region c) in Lemma 2..1) for which a strictly positive amount of time is allocated to both tasks, the static comparative of $v^*(a)$ and $q^*(a)$ is the one summarized in Table 2.

	α	β	δ	t	a
$v^*(a)$	$+$	$-$	$+$ if and only if $a^2\left(\beta^2 + t^2\delta^2\right) > \alpha^2$	$+$	$+$ if and only if $a^2\left(t^2\delta^2 - \beta^2\right) > \alpha^2$
$q^*(a)$	$-$	$+$	$+$ if and only if $\alpha > a\beta$	$+$	$+$

Table 2

As shown in Table 2, the value of the project $v^*(a)$ is increasing in α and decreasing in β. It only increases in δ for high levels of a, i.e, for $a^2 > \frac{\alpha^2}{\beta^2 + t^2\delta^2}$. This threshold decreases with the complementarity among senior and junior scientists in the project. Also, $v^*(a)$ is decreasing in a when $t\delta < \beta$. When $t\delta \geq \beta$, it decreases in a if a is smaller than a certain cut-off, $a > \frac{\alpha}{\sqrt{(\beta+\delta t)(\delta t - \beta)}}$ and increases otherwise. We finally note that $v^*(a)$ is convex in a. As to $q^*(a)$, we can see that the expected capability is linear in a and the static comparative provides us the predicted intuitive results.

We would now consider different projects defined by the pair (α, δ). For research projects that differ on α, those projects whose success is heavily based on the time the senior scientist directly spends on them, will have a better final quality. This finding is true because the senior scientist will dedicate more of her time to the project,

allocating more effort to research and leaving junior scientist's guidance, and hence his final capabilities are at low levels. If we compare two research projects that differ only by δ, there are cases where in the project with higher δ, the senior scientist can produce better innovative results and more capable junior scientists. Senior researchers that have projects that differ in both dimensions (α, δ) are more difficult to compare because the effects go in opposite dimensions.

Ex-post Project Value and the Junior Scientist Final Capability After focusing on the ex-ante equilibrium values, we now compute the *ex-post* project value and the *ex-post* junior scientist's capability that are obtained in equilibrium. We use the case of perfect information about the innate ability of the junior as a benchmark to measure the difference between ex-ante and ex-post efficiency under ignorance.[11] When the senior scientist decides on the time allocation of efforts under ignorance, the allocation of time depends on her beliefs $a = E(a)$. However, the ex-post project value (resp., the junior capability) depends on the true innate ability of the junior scientist $\hat{a}$. Hence, we denote the ex-post levels as $v^o(\hat{a}, E(a))$ and $q^o(\hat{a}, E(a))$.

As a function of the parameters, the ex-post outcomes are:

a) When $\beta > \delta t$ and $a > \frac{\alpha}{\beta - \delta t}$,

$$v^o(\hat{a}, E(a)) = 0 \text{ and } q^o(\hat{a}, E(a)) = \hat{a}t$$

b) When $a < \frac{\alpha}{\beta + \delta t}$,

$$v^o(\hat{a}, E(a)) = \alpha t \text{ and } q^o(\hat{a}, E(a)) = 0$$

[11] With complete information about the junior scientist's innate ability $\hat{a} = E(a)$, and the ex-post and ex-ante values of the project (resp., the junior capability) coincide. The corresponding expression is obtained by substituting a for $\hat{a}$ in the expression of the above mentioned subsection.

c) Otherwise, using $x \equiv \frac{E(a)(\delta t-\beta)+\alpha}{2E(a)\delta}$ and $y \equiv \frac{E(a)(\delta t+\beta)-\alpha}{2E(a)}$,

$$v^o(\hat{a}, E(a)) = v^o(\hat{a}, E(a)) = \alpha x + \hat{a}xy \text{ and } q^o(\hat{a}, E(a)) = \hat{a}\left(\frac{t}{2} - \frac{\alpha - \beta E(a)}{2\delta E(a)}\right)$$

We comment first on the ex-post value of the project. In the interior case, $v^o(\hat{a}, E(a))$ is linear and increasing in $\hat{a}$.[12] To measure the distortion ex-post on the value of the project, i.e., we compute $v^*(\hat{a})$ and $v^o(\hat{a}, E(a))$ and depict the difference. The ex-post value of the project may be higher or lower under ignorance than under full information. Intuitively, under ignorance the value will be higher when expectations on the innate ability of the junior are high and its true level low. Figure 2..2 represents the comparison for different levels of $E(a)$ when the concern for training is low enough, $\delta t \geq \beta$. Under ignorance, the project value is higher than under perfect information for low levels of the true innate ability and the opposite occurs when it is high.[13]

In the case where there is a high enough concern for training, $\delta t < \beta$, this distortion is more extreme. As Figure 2..3 shows, the value under ignorance keeps increasing (dotted lines) but the value under full information is always non-increasing in $\hat{a}$ (full lines). Also, for intermediate values of the expected innate ability, the senior scientist is always overinvesting in the training task.

A more interesting situation is to analyze how the ex-post project value under imperfect information changes with the senior scientist's prior belief about the ability

[12] Because $\text{sign}(xy) = \text{sign}(E(a)(\delta t - \beta) + \alpha))(E(a)(\delta t + \beta) - \alpha)$ is positive in this region.

[13] Note that this comparison is between the decision of the senior scientist with full information and the decision under ignorance, but none of these decisions may be in accordance with the social optimum. The distortion at the bottom that may lead senior scientists to train more juniors under ignorance may be good from a social point of view. The distortion at the top may indicate more of a concern for a society interested in guaranteeing that excellent projects and high-potential junior scientists are better identified. We discuss these issues in Section 5.

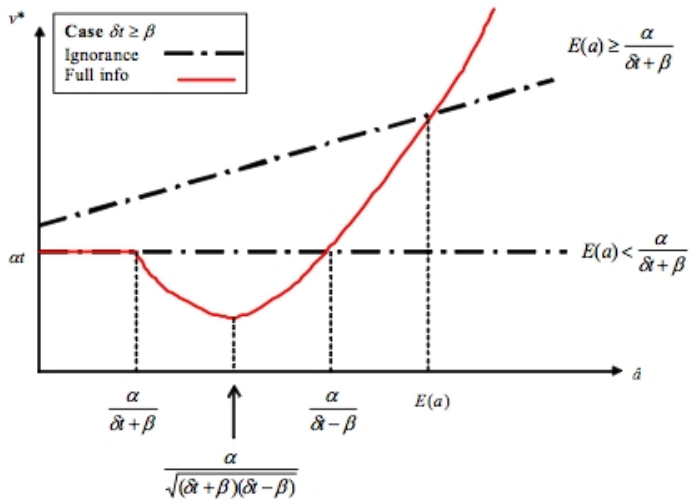

Figure 2..2: $v^*(\hat{a})$ and $v^o(\hat{a}, E(a))$ when $\delta t \geq \beta$

of the junior scientist, $E(a)$. Given $\hat{a}$, when $E(a)$ is low enough,[14] an increase in $E(a)$ leads to an increase in $v^o(\hat{a}, E(a))$. Otherwise it will decrease. One can also read this result the other way around. Given $E(a)$, for high values of $\hat{a}$ a higher prior will increase $v^o(\hat{a}, E(a))$, and for low $\hat{a}$ a higher prior will decrease $v^o(\hat{a}, E(a))$. This means that a higher prior in the case of Figure 2..2, increases the distortion between $v^*(\hat{a})$ and $v^o(\hat{a}, E(a))$ for intermediate values of $\hat{a}$ but decreases the distortion for very low and very high values of $\hat{a}$. So, when the junior scientist is indeed very good, this inefficiency becomes smaller as $E(a)$ is closer to $\hat{a}$. In the case of $\delta t < \beta$ (Figure 2..3), the effect of an increase in $E(a)$ will decrease the distortion for low levels of $\hat{a}$ and increase it onwards.

We now consider the distortions of the junior scientist's final capability. Note that for the interior solution, $q^o(\hat{a}, E(a))$ is linear and increasing in $\hat{a}$ and increasing

[14] For $E(a) \leq \left(\frac{\hat{a}\alpha}{2(\alpha+\hat{a}\beta)}\right)^{\frac{1}{2}}$.

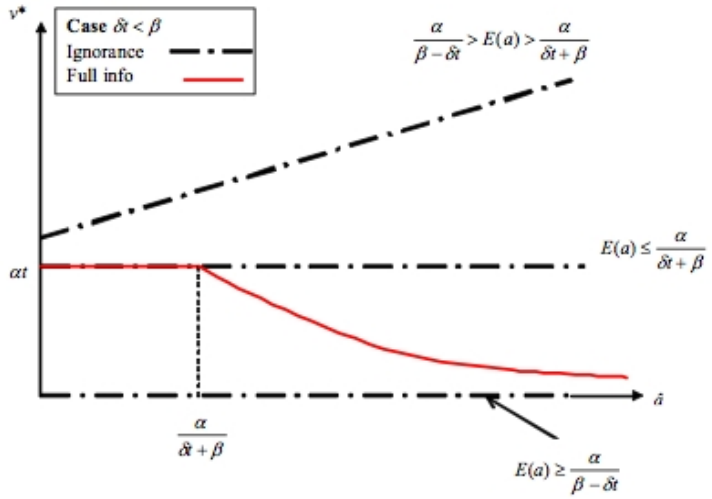

Figure 2..3: $v^*(\hat{a})$ and $v^o(\hat{a}, E(a))$ when $\delta t < \beta$

and concave in $E(a)$. Comparing the ex-post junior's training under full information $q^*(\hat{a})$ and under ignorance $q^o(\hat{a}, E(a))$, we obtain the results depicted in Figures 2..4 and 2..5. Figure 2..4 represents the case when the concern for training is low, $\delta t \geq \beta$. We can see from the two curves that there is over-training of low-ability and under-training of high-ability juniors. The further away the senior scientist's prior is from the true ability of the junior, either from above or below, the higher is the distortion in the ex-post formation, and the distortion increases proportionately. The effect of a higher prior $E(a)$ over the junior scientist's ability is a higher slope of the "ignorance" ex-post curve, which means that the distortion increases for lower values of $\hat{a}$ and will decrease for higher values. More accurate training is provided to the population of junior scientists with more potential.

For $\delta t < \beta$, as illustrated in Figure 2..5, the distortion and the effects of a higher $E(a)$ are similar.

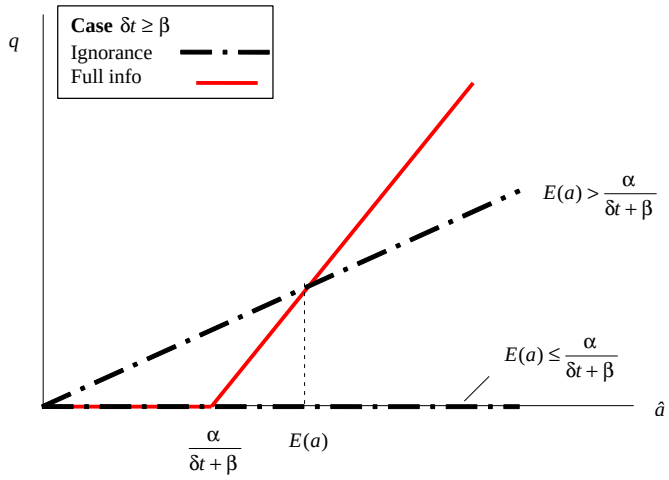

Figure 2..4: $q^*(\hat{a})$ and $q^o(\hat{a}, E(a))$ when $\delta t \geq \beta$

As in the case with the value of the project, the sign of the distortion may be aligned or not aligned with social interest. We will discuss this aspect in Section 2.5. Note also that if the interval $[\underline{a}, \bar{a}]$ is smaller or if, for a given interval the expected innate ability of the junior scientist, $E(a)$ is higher, then it may be the case that more education is provided under ignorance and the distortion is not too big.

2.4 Extensions

Let us consider two natural questions that may come to mind that we present here in two independent extensions. In the first one, we allow for the senior scientist to choose the amount of time that she will work. In other words, t is not exogenous but her choice. In the second extension we assume that the time available is exogenously given, but we allow the senior scientist to have access to a better pool of junior scientists at the cost of some of her time.

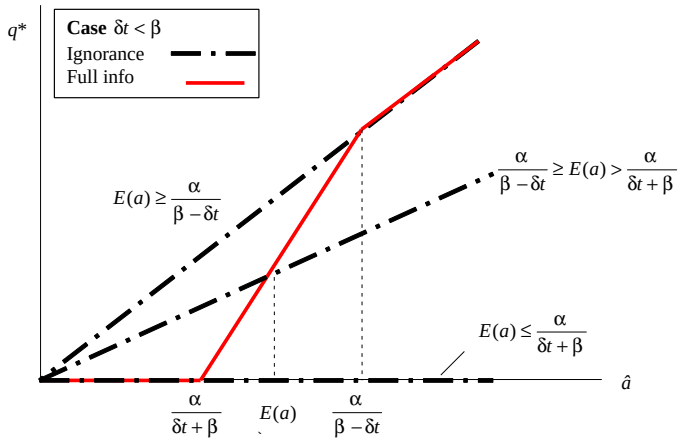

Figure 2..5: $q^*(\hat{a})$ and $q^o(\hat{a}, E(a))$ when $\delta t < \beta$

Both extensions can be viewed as a sequential decision problem where, once either the time allocated to work or the ability of the junior scientist is determined, the analysis of the previous sections tells us the result in terms of research and training. Hence, it is useful to use the optimal allocation of time as a function of t and a to write the utility of the senior in terms of these variables. Using Lemma 2..1 we see that:

a) When $\beta > \delta t$ and $a > \frac{\alpha}{\beta - \delta t}$,

$$u^*(a,t) = \beta a t$$

b) When $a < \frac{\alpha}{\beta + \delta t}$,

$$u^*(a) = \alpha t$$

c) Otherwise,

$$u^*(a,t) = \frac{(a\delta t + \alpha)^2 - (\beta a)^2}{4a\delta} + \beta\left(\frac{ta}{2} - \frac{\alpha - \beta a}{2\delta}\right)$$

Total Time Worked Until now we have considered that the time t the senior scientist works is fixed. Let us now consider that, as in the line of more traditional moral hazard models, the senior scientist may decide the total amount of time t she will devote to work (the total effort). To determine this total working time t, the senior scientist maximizes her expected utility net of the cost of the working time. We will assume that the cost of working time c is high enough. More precisely, we assume $c \geq \frac{a\delta}{2}$.[15] Hence, the senior solves

$$\underset{t}{Max} \left\{ u^*(a,t) - \frac{c}{2}t^2 \right\}.$$

From this problem, we obtain the following result:[16]

Lemma 2..2 *The senior scientist's total time as a function of the parameters is*

a) When $\alpha - a\beta \geq 0$ *and* $c \geq \frac{a\delta\alpha}{\alpha - a\beta}$,

$$t^* = \frac{\alpha}{c}$$

b) When $\alpha - a\beta \geq 0$ *and* $\frac{a\delta}{2} < c \leq \frac{a\delta\alpha}{\alpha - a\beta}$ *or* $\alpha - a\beta < 0$ *and* $\frac{\beta a^2\delta}{a\beta - \alpha} > c$,

$$t^* = \frac{\alpha + a\beta}{2c - a\delta}$$

c) When $\alpha - a\beta < 0$ *and* $\frac{a\delta}{2} < c \leq \frac{\beta a^2\delta}{a\beta - \alpha}$,

$$t^* = \frac{a\beta}{c}$$

[15] If the cost c is smaller that $a\delta/2$ the optimal time goes to infinite. In this case it would be natural to include a maximum time limit T. We will comment on this assumption later, but we will concentrate on the case where c is high for the sake of simplicity.

[16] Note that for $\alpha - a\beta \geq 0$, we have $\frac{a\delta}{2} \leq \frac{a\delta\alpha}{\alpha - a\beta}$, and for $\alpha - a\beta < 0$, we have $\frac{a\delta}{2} < \frac{\beta a^2\delta}{a\beta - \alpha}$. Hence, the regions of Lemma 2..2 are well defined.

Lemma 2..2 is depicted in Figure 2..6 in the space $(c, (\alpha - a\beta))$.

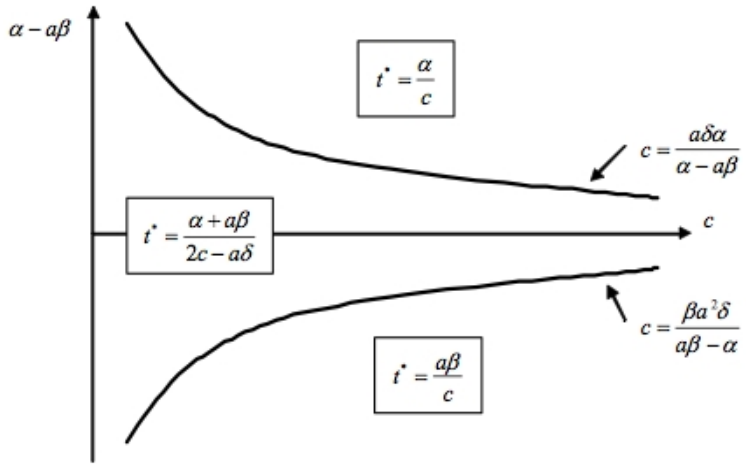

Figure 2..6: Optimal t^* in the space $(c, (\alpha - a\beta))$

From Lemma 2..2 we conclude that, as expected, the time the senior scientist works t^* is a non-decreasing function of a, α, β, δ and is decreasing in c. The comparative statics are summarized in Table 3.

	α	β	δ	c	a
$t^* = \frac{\alpha}{c}$	$+$	0	0	$-$	0
$t^* = \frac{\alpha+a\beta}{2c-a\delta}$	$+$	$+$	$+$	$-$	$+$
$t^* = \frac{a\beta}{c}$	0	$+$	0	$-$	$+$

Table 3

Given the optimal t^*, we compute the time allocated to each task. As expected, the time allocated to both tasks is decreasing in c (that now plays a role similar to

a decrease in t in the previous sections). Effort e_R is non-increasing and e_G is non-decreasing in the concern for training, β. More precisely, for $t^* = \frac{\alpha}{c}$, the optimal allocation of time is $\left(e_R = \frac{\alpha}{c}, e_G = 0\right)$ and research increases with α but nothing is affected by β or a. At the other extreme, for $t^* = \frac{a\beta}{c}$ the optimal allocation of time is $\left(e_R = 0, e_G = \frac{a\beta}{c}\right)$. For the case $t^* = \frac{\alpha + a\beta}{2c - a\delta}$ the time allocated to both tasks deserves some attention, and we perform comparative statics given in Corollary 2..2.

Corollary 2..2 *When $t^* = \frac{\alpha + a\beta}{2c - a\delta}$, region b) in Lemma 2..2, we have that the allocation of time to the tasks is*

$$\begin{aligned} e_R^* &= \frac{1}{2}\left(\frac{\alpha + a\beta}{2c - a\delta} + \frac{\alpha - a\beta}{a\delta}\right) \\ e_G^* &= \frac{1}{2}\left(\frac{\alpha + a\beta}{2c - a\delta} - \frac{\alpha - a\beta}{a\delta}\right) \end{aligned}$$

and the comparative statics are presented in Table 4:

	α	β	a	δ
$e_R^*(a, \alpha, \beta, \delta)$	$+$	$+$ *iff* $c < a\delta$	$+$ *iff* $c < \frac{\delta a}{2}\frac{(2\alpha + a\beta)}{\alpha}$	$+$ *iff* $c < \frac{\delta a}{2}\left(\frac{(\alpha + a\beta)^{1/2}}{(\alpha - a\beta)^{1/2}} - 1\right)$
$e_G^*(a, \alpha, \beta, \delta)$	$+$ *iff* $c < a\delta$	$+$	$+$	$+$

Table 4

For completeness let us remark that in a version of the model where c is not constrained from below, and there is a maximum amount of time T that the senior scientist has available, the results will be similar, except for low costs c $\left(c \leq \frac{a\delta}{2}\right)$ and/or T small enough $\left(T < \frac{|\alpha - a\beta|}{a\delta}\right)$. In these cases, the senior chooses to work for all the

available time and she allocates all the time T either to research ($\alpha > a\beta$) or to training ($\alpha < a\beta$) (except if $\alpha = a\beta$, case where she is indifferent). Changes in α or in β do not affect the total time allocated to work and only discrete changes may affect to which task this time T is allocated. In these cases, only changes of the time available for these activities may have an effect on the senior scientist's behavior.

When Expected Ability and Time Are Related In Section 2.2 we analyzed the decisions of a senior scientist that is (randomly) matched with a junior scientist of innate ability a. However, one may wonder about what happens if the junior's expected innate ability a depends on some previous activity that the senior performs and that consumes time. This may correspond to a selection process that tries to identify a better population of junior scientists, an advertising or investment procedure that aims to attract a junior scientist with a higher expected innate ability, or an undergraduate system that provides better skills and better information about the juniors' abilities. Here, we model the relationship between the time invested in increasing the innate ability of the junior she works with and the remaining time available for research and training.

Let us assume that T is the maximum amount of time available for the three tasks and A the innate ability of the junior scientist if no effort is made to improve it. Let us denote by $g(T-t)$, with $g > 0$, the improvement of the innate ability of the junior scientist that the senior can obtain by using an amount of time $(T-t)$ to improve the quality of the junior with whom she works, in such a way that she will have t to allocate to the tasks of research and formation. Hence, the ability of the junior scientist with whom the senior will work with is $a = A + g(T-t)$.

In this case, the senior scientist chooses (a, t) by maximizing her utility function, taking into account the constraints $a = A + g(T - t)$. To simplify presentation and to avoid cumbersome calculations for different regions of parameters, we just present an example where we assume that the senior has no appreciation of the junior scientist's final capability ($\beta = 0$) and that the complementarity effect is 1 ($\delta = 1$). This implies that we will be in Region $\delta t \geq \beta$ (that, in this case, is reduced to $t \geq 0$). Under this parameter combination, as a function of t, the senior's allocation of time depends on whether $a \geq \frac{\alpha}{\delta t}$ (and her time will be allocated to research and formation) or $a \leq \frac{\alpha}{\delta t}$ (and she will only do research).

Lemma 2..3 *Assuming $\beta = 0$, $\delta = 1$ and the relation between time and ability given as $a = A + g(T - t)$, the senior scientist's decision on the optimal innate ability and on the optimal amount of time spent in previous activities is:*

a) When $(gT + A)^2 - 12\alpha g \geq 0$,

$$a^* = \frac{gT + A + \sqrt{(gT + A)^2 - 12\alpha g}}{6} \text{ and } t^* = \frac{5(gT + A) - \sqrt{(gT + A)^2 - 12\alpha g}}{6g}$$

b) When $(gT + A)^2 - 12\alpha g < 0$,

$$a^* = A \text{ and } t^* = T$$

Lemma 2..3 illustrates that projects in which the senior's direct research productivity (α) is low enough, she is willing to spend time in selection activities in order to work with a junior scientist of higher expected innate ability. This finding is intuitive since, being a less productive scientist, she wants to increase the prospects of working

with a more talented junior scientist. When α is high enough, then she may choose not to spend any time in activities to obtain more information about the junior's innate ability, leaving it at level A. This way, she chooses to allocate all of the time resources to training and research only. Note also that for a given combination of the other parameters (g, α), when T or A are small it is more often the case that the senior's optimal decision is not spend time in improving the innate ability of the junior (while this does not mean that she will not allocate some time to formation).

2.5 Welfare Analysis

We would like to consider here a situation with a social planner who is concerned about the level of research that the senior scientist achieves and the final capacity of the junior, that he interprets as a measure of the potential of the next generation of researchers. We consider first that this social planner has the welfare function:

$$W = E(v^o(\hat{a}, E(a))) + \lambda E(q^o(\hat{a}, E(a))),$$

where λ can be interpreted as the society's relative concern about the capability of the next generation of researchers.

If λ coincides with β, then the decision of the senior and the aims of the society concur. If λ and β do not coincide, the social planner may be tempted to intervene. To discuss this possibility, we take as a starting point our basic model presented in Section 2, where we assume that there is no moral hazard problem, just a decision about the allocation of time. An alternative way of looking at the comparative static in Table 1 (and the discussion after it) is to consider how society may induce changes

in some parameters to affect the senior scientist's allocation of time to research and formation. For this purpose, the social planner must affect the senior's utility $u = \alpha e_R + \delta a e_G e_R + \beta a e_G$, possibly using (α, δ) and β as instruments.[17]

If the outcome of research v and the outcome of training q are verifiable, the regulator can manipulate the decision of the senior scientist by changing her awareness about these two variables. The planner can increase the senior scientist's utility from the project's value (that is, increasing α and δ in the same proportion or, equivalently decreasing β) or to increase the senior's payoff as a function of the quality of the junior she mentors (increasing β) via the definition of a successful career or the allocation of research funds that weigh this aspect of the academic career. Note that increasing both perceptions is useless when the aim is to change the allocation of total time, because total time is fixed. Also, if only publications (and other measures of the senior scientist project results) are verifiable, the social planner can only encourage more time to research (through tenure tack rules, opportunities to travel and access to research funds, or peer esteem, which in our model corresponds to decrease β) but he cannot increase it above the natural inclination of the senior scientist. Only by discouraging research can the time allocated to training be increased.

The social planner can change the junior scientist's innate quality a (for example, by having an attractive and selective program of fellowships) that allows the attraction of better students. Indeed, several European expert institutions (e.g., EURAB, ESF) have given priority to the training of scientists and developed actions so that postdoc-

[17] Obviously, the social planner can also change the time available for these tasks t (for example, by reducing the senior's involvement in other time-consuming tasks, such as administrative ones).

toral researchers ascend to PIs in recent years. These actions involve providing access to special grants, as well as promoting free and secure mobility.

When total time is fixed, these instruments have a positive effect on one task but a negative effect on the other. Both efforts only increase simultaneously by inducing a higher t, as already mentioned. If a moral hazard situation exists, and the senior scientist decides how much time to work, the previous discussion of the instruments to use holds partially. In this case, incentivizing the results for both research and training may be optimal because these instruments affect not only time allocation but also how much time the senior decides to work. As shown in Corollary 2..2, if the cost of the effort or the quality of the junior is high enough, then increases in (α, δ) , which correspond to a higher utility associated to the value of the research project, or increases in β induce more research and more training (because they induce more incentives to work). If juniors are gifted enough, both instruments (increasing the utility the senior scientist receives from research or from training) have positive effects on the senior scientist's dedication to both tasks. In a society where the population of juniors is of low expected ability, the instruments have positive effects on one task and negative on the other and encouraging one activity crowds out the effort on the other one. This emphasizes the importance of attracting a good population of junior scientists. This comment connects with the analysis conducted in Section 2.4 where Lemma 2..3 draws attention to the possibility that the population of juniors can be linked to the time allocated to select them. Note however, that measures that increase a^* will decreases t^*. This may lead to an increase in the senior's dedication to a task but may trigger a decrease in the time allocate to the other task unless the cost of obtaining better pools of junior scientists,

g, decreases.

Another point of view is to consider that the social planner is not just concerned about the expected level of research or training, but about reaching high enough outcomes in both tasks. In other words, it can be the case that the social planner is only interested in excellence and, hence, in achieving both the highest quality in innovation and the highest level of ex-post capability.[18] Imagine a social planner considers research to be valuable only if $v \geq V$ and wants junior scientists to be endowed with a minimum final capability $q \geq Q$ to be considered good independent researchers. In this framework, the social planner cares about $\tilde{W}$,

$$\tilde{W} = E_{v \geq V}(v^o(\hat{a}, E(a))) + \lambda E_{q \geq Q}(q^o(\hat{a}, E(a))),$$

where the minimum requirements (V, Q) are given by the social planner a level of exigency.

We use now the results presented in Section 2.3 based on the model where time t is given. In order to have projects and young researchers above the cutoffs (V, Q) with high probability (or a high proportion), the social planner can use the available instruments β, a, t. We have seen in Figures 5 and 6 that the level of final capability under ignorance, $q^o(\hat{a}, E(a))$, increases with the senior's priors with respect to the innate ability of the junior.

To help the discussion along, in Figure 2..7 (using the information conveyed in

[18] The reason may be that the society may not consider results below a minimum requirement on both outcomes as an achievement: the society may value only "good enough" discoveries to be patented or to improve knowledge, and only "capable enough" junior scientists may be considered good researchers.

Table 2) we represent in the space (a, t) the expected value of the project in equilibrium, as well as the iso-project value curves and the iso-final capability curves of the junior scientist, keeping constant other parameters (the dotted lines). This figure shows that a higher total amount of time always induces a higher expected project value as well as a higher junior scientist capability. However, increasing only a does not have the same effect. If the social planner wants project values to have at least value V' and the junior capability Q' then it has, on one hand, to procure a higher total time t available to the senior scientist, and on the other hand induce as much as possible a selected junior scientist with enough potential.

For the social planner, the senior's priors is a possible instrument to obtain a superior outcome in the training component because a higher prior induces more time allocated to training. Besides the quantity effect, which is the fact that more of the population reaches an independent research status (attains q above Q), there is a quality effect on junior researchers since they are better prepared. Analyzing v^o here, we conclude that the project value under ignorance increases with the senior scientist's priors, but only until a certain point. For very high values of the expected ability the equilibrium value of the project starts to decrease. Hence, when fixing the level of $E(a)$ both effects must be taken into account. Increasing the expected ability of the juniors population, $E(a)$, can be performed by implementing or increasing subsidies to a tougher selection of scientists eligible to perform research under supervision.[19] Im-

[19] One can assume that the senior sceintist or ither department are in a better position to assess the ability of a junior scientist, but it also seems reasonable to think that higher resources allocated to the selection processes may help. Any selection process would include the past education of the junior sceintist, as well as considering the university of origin and inviting the juniors for an interview. If these resources are not available, their own students may be less risky that outsiders.

plementing good programs in earlier education can also cause this shift. Also, offering more attractive conditions in programs for PhD and postdocs may attract better candidates for the task, who are otherwise drawn to more attractive careers in other sectors. These conditions may be better stipends or better lab equipment accessible to junior scientists. Implementing such measures can shift the population to higher levels of innate ability, first order stochastically dominating the initial population distribution or even increase the lower bound $\underline{a}$ of the distribution on the innate ability of junior scientists (which can correspond to a higher A in the analysis of Section 2.4).

The productivity of the senior scientist with respect to her own research, α, can also be used to either promote higher project value, and the concern that the senior has over formation, β, to promote the higher final capability of junior scientists. The social planner will obviously face a trade-off here as well. However, some outcomes are unreachable controlling only the parameter β; if the regulator wants to have simultaneously a high enough expected project value (above V) and a high enough expected q^* (above Q), this objective may not be reached only by manipulating β. If these cutoffs V and Q are very demanding, they can be obtained only for some combination of the remaining parameters (a, t, α, δ).

2.6 Conclusion

In this paper we provide a model where not only research results of scientists are important, but also their involvement and effect in training young scientists. To this aim we propose a multitask model where the training of a junior scientist depends on the incentives that a senior scientist has to allocate time to training when this is

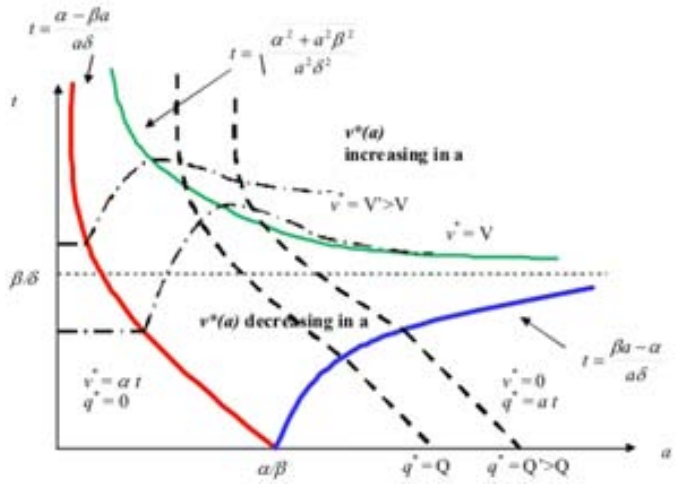

Figure 2..7: v^* and q^* as function of (a, t)

at the cost of spending less time doing research. Our model proposes the (testable) predictions that researchers are more inclined to be involved in training the more time they have to allocate to these tasks, the higher the expected innate ability of the junior scientist, and the less important is her own involvement for the success of the research project is as compared to the complementarities of working with a competent junior.

We study the possibility that a senior engages in previous activities to obtain more information about the innate ability of the junior. This possibility is used when her own productivity on the project is not too high and investing in this activity provides better chances of working with a good junior. We also analyze the distortions that arise in the value of a research project and in the final capability of the junior scientist between imperfect information (ex-ante efficiency) and full information (ex-post efficiency). The patterns of the distortions will vary with the relevant parameters, and more or less training may be provided when we compare the decision of the senior under ignorance or full information. This aspect is also important when a regulator consid-

ers policies to induce more training and decrease these distortions. We also discuss that, in some cases, if a regulator implements attractive, recognized training programs in earlier education (leading to a higher proportion of good junior scientists), as well as a tougher selection process for research under supervision and attractive conditions to appointments, the proportion and the quality level of independent scientists in the population will be positively affected.

This model is still a first approach to the problem. We think that it may help to provide a simple framework to conduct some empirical studies and to think about policy issues. This model has the potential to be generalized toward the study of additional problems where the matching among a senior scientist and junior scientist, the senior's choice of a research project, or the team component of some research processes are included.

2.7 Appendix

Proof of Lemma 2..1 The program can be rewritten as

$$\underset{e_R,e_G}{Max}\ \alpha e_R + a\delta\left(t-e_R\right)e_R + \beta a\left(t-e_R\right)$$

$$s.t. \qquad e_R \geq 0 \qquad \text{and} \qquad e_R \leq t$$

The Lagrangian of this program is $\mathcal{L} = \alpha e_R + a\delta\left(t-e_R\right)e_R + \beta a\left(t-e_R\right) + \lambda e_R + \mu\left(t-e_R\right)$. Its FOC is

$$\alpha + a\delta\left(t-2e_R\right) - \beta a + \lambda - \mu = 0.$$

Then the possibilities are (1) $\lambda = 0$, $\mu = 0$ and $e_R = \frac{a\delta t+\alpha-\beta a}{2a\delta}$, which is a candidate when $\frac{a\delta t+\alpha-\beta a}{2a\delta} \geq 0$ and $\frac{a\delta t+\alpha-\beta a}{2a\delta} \leq t$, or equivalent, when $\alpha+a\delta t-\beta a \geq 0$ and $\alpha-a\delta t-\beta a \leq 0$, (2) $\lambda > 0$, $\mu = 0$ and $e_R = 0$, which is a candidate when $\alpha + a\delta t - \beta a \leq 0$ (3) $\lambda = 0$, $\mu > 0$ and $e_R = t$, which asks for $\alpha - a\delta t - \beta a \geq 0$. Finally, the combination of these cases and the definition of $e_G = t - e_R$ give the results

Solution e_R^*	Solution e_G	When $\delta t \geq \beta$	When $\delta t < \beta$
$e_R^* = 0$	$e_G = t$		$\frac{\alpha}{\beta-\delta t} < a$
$e_R^* = \frac{t}{2} + \frac{\alpha-\beta a}{2a\delta}$	$e_G = \frac{t}{2} - \frac{\alpha-\beta a}{2a\delta}$	$\frac{\alpha}{\beta+\delta t} \leq a$	$\frac{\alpha}{\beta+\delta t} \leq a \leq \frac{\alpha}{\beta-\delta t}$
$e_R^* = t$	$e_G = 0$	$a < \frac{\alpha}{\beta+\delta t}$	$a < \frac{\alpha}{\beta+\delta t}$

presented in the Lemma.

Proof of Lemma 2..2 We proceed for simplicity by regions of parameters:

Case 1: $\alpha - a\beta > 0$

In this case:

$$u(t) = \frac{(a\delta t+\alpha)^2-(\beta a)^2}{4a\delta}+\frac{2\beta a\,(a\delta t-\alpha+\beta a)}{4a\delta} \text{ if } t \geq \frac{\alpha-a\beta}{a\delta} \quad (2..3)$$

$$u(t) = \alpha t \text{ if } t \leq \frac{\alpha-a\beta}{a\delta} \quad (2..4)$$

Step (i) Let's first consider the function $u(t)=\alpha t$. The maximization problem is

$$\underset{t}{Max}\ \left\{\alpha t-\frac{c}{2}t^2\right\}$$
$$s.t. \qquad 0 \leq t \leq \frac{\alpha-a\beta}{a\delta}$$

The Lagrangian function is $\mathcal{L}=\alpha t-\frac{c}{2}t^2+\lambda(\frac{\alpha-a\beta}{a\delta}-t)+\mu t$. The FOC is

$$\alpha-ct-\lambda+\mu=0.$$

Then the possibilities are (i.1) $\lambda=0$ and $\mu=0$, then $t=\frac{\alpha}{c}$, which is a candidate when $\frac{\alpha}{c}\leq\frac{\alpha-a\beta}{a\delta}\Longleftrightarrow\frac{a\delta\alpha}{\alpha-a\beta}\leq c$. (ii.2) $\lambda>0$ and $\mu=0$, then $t=\frac{\alpha-a\beta}{a\delta}$, which is a candidate when $\alpha-c(\frac{\alpha-a\beta}{a\delta})>0$, i.e., $\frac{\alpha}{c}>\frac{\alpha-a\beta}{a\delta}\Longleftrightarrow\frac{a\delta\alpha}{\alpha-a\beta}>c$. (ii.3) $\lambda=0$ and $\mu>0$, then $t=0$, which is a candidate if $\alpha<0$ which is never the case.

Step (ii) Let's now consider the function $u(t)=\frac{(\alpha+at\delta)^2-(a\beta)^2}{4a\delta}+\frac{2a\beta(a\delta t-\alpha+a\beta)}{4a\delta}$. The maximization problem is

$$\underset{t}{Max}\ \left\{\frac{(\alpha+at\delta)^2-(a\beta)^2}{4a\delta}+\frac{2a\beta(a\delta t-\alpha+a\beta)}{4a\delta}-\frac{c}{2}t^2\right\}$$
$$s.t. \qquad t \geq \frac{\alpha-a\beta}{a\delta}$$

The Lagrangian is $\mathcal{L}=\frac{(\alpha+at\delta)^2-(a\beta)^2}{4a\delta}+\frac{2a\beta(a\delta t-\alpha+a\beta)}{4a\delta}-\frac{c}{2}t^2+\lambda(t-\frac{\alpha-a\beta}{a\delta})$. The FOC is

$$\frac{\alpha+at\delta+a\beta}{2}-ct+\lambda=0.$$

Then, the possibilities are: (ii.1) $\lambda=0$ and $t=\frac{\alpha+a\beta}{2c-a\delta}$, which is a candidate when $\frac{\alpha+a\beta}{2c-a\delta}>\frac{\alpha-a\beta}{a\delta}$ or, equivalently, when $\frac{a\delta\alpha}{\alpha-a\beta}\geq c$. (ii.2) $\lambda>0$ and $t=\frac{\alpha-a\beta}{a\delta}$, which asks for

$\frac{\alpha+a\delta(\frac{\alpha-a\beta}{a\delta})}{2} - c(\frac{\alpha-a\beta}{a\delta}) < 0$, or equivalently $\frac{a\delta\alpha}{\alpha-a\beta} < c$.

Step (iii) Comparing the results from the previews steps, and realizing that the candidate for solution in one of cases is a possible solution in the other case we obtain the result presented in the Lemma.

Case 2: $\alpha - a\beta < 0$

Here we have:

$$u(t) = \frac{(a\delta t+\alpha)^2 - (\beta a)^2}{4a\delta} + \frac{2\beta a\,(a\delta t - \alpha + \beta a)}{4a\delta} \text{ if } t \geq \frac{a\beta-\alpha}{a\delta} \quad (2..5)$$

$$u(t) = a\beta t \text{ if } t \leq \frac{a\beta-\alpha}{a\delta} \quad (2..6)$$

Step (i) Let's consider the function $u(t) = a\beta t$. The maximization problem is

$$\underset{e_R,e_G}{Max} \left\{a\beta t - \frac{c}{2}t^2\right\}$$
$$s.t. \qquad 0 \leq t \leq \frac{a\beta-\alpha}{a\delta}$$

The Lagrangian function is $\mathcal{L} = a\beta t - \frac{c}{2}t^2 + \lambda(\frac{a\beta-\alpha}{a\delta} - t) + \mu t$. The FOC is

$$a\beta - ct - \lambda + \mu = 0.$$

Then the possibilities are (i.1) $\lambda = 0$ and $\mu = 0$ which gives $t = \frac{a\beta}{c}$, which is a candidate when $\frac{a\beta}{c} \leq \frac{a\beta-\alpha}{a\delta} \iff \frac{\beta a^2\delta}{a\beta-\alpha} \leq c$. (ii.2) $\lambda > 0$ and $\mu = 0$, and $t = \frac{a\beta-\alpha}{a\delta}$, which is a candidate when $\alpha - c\left(\frac{a\beta-\alpha}{a\delta}\right) > 0$, i.e., $\frac{\alpha}{c} > \frac{a\beta-\alpha}{a\delta} \iff \frac{\beta a^2\delta}{a\beta-\alpha} > c$. (ii.2) $\lambda = 0$ and $\mu > 0$, and $t = 0$, which is a candidate when $a\beta < 0$, which is never the case.

Step (ii) Let's now consider the function $u(t) = \frac{(\alpha+at\delta)^2-(a\beta)^2}{4a\delta} + \frac{2a\beta(a\delta t-\alpha+a\beta)}{4a\delta}$. The max-

imization problem is

$$\underset{t}{Max}\ \left\{ \frac{(\alpha + at\delta)^2 - (a\beta)^2}{4a\delta} + \frac{2a\beta(a\delta t - \alpha + a\beta)}{4a\delta} - \frac{c}{2}t^2 \right\}$$

$$s.t. \qquad t \geq \frac{a\beta - \alpha}{a\delta}$$

The Lagrangian function is $\mathcal{L} = \frac{(\alpha+at\delta)^2-(a\beta)^2}{4a\delta} + \frac{2a\beta(a\delta t-\alpha+a\beta)}{4a\delta} - \frac{c}{2}t^2 + \lambda(t - \frac{a\beta-\alpha}{a\delta})$. The FOC is

$$\frac{\alpha + at\delta + a\beta}{2} - ct + \lambda = 0.$$

Then, the possibilities are: (ii.1) $\lambda = 0$ and $t = \frac{\alpha+a\beta}{2c-a\delta}$, which is a candidate when $\frac{\alpha+a\beta}{2c-a\delta} > \frac{a\beta-\alpha}{a\delta}$ or, equivalently, when $\frac{\beta a^2\delta}{a\beta-\alpha} > c$. Note that $\frac{\beta a^2\delta}{a\beta-\alpha} < \frac{a\delta}{2}$ (ii.2) $\lambda > 0$ and $t = \frac{a\beta-\alpha}{a\delta}$, which asks for $\frac{\alpha+a\delta(-\frac{\alpha-a\beta}{a\delta})+a\beta}{2} - c(\frac{a\beta-\alpha}{a\delta}) < 0$, or equivalently $\frac{\beta a^2\delta}{a\beta-\alpha} < c$.

Step (iii). Comparing the results from the previews steps, and realizing that the candidate for solution in one of cases is a possible solution in the other case we obtain the result presented in the Lemma.

Case 3: $\alpha - a\beta = 0$

In this case:

$$u(t) = t\frac{2(\alpha + \beta a) + a\delta t}{4} \text{if } t \geq 0$$

Then the maximization problem of the senior is:

$$\underset{t}{Max}\ \left\{ \frac{2t(\alpha + \beta a) + a\delta t^2}{4} - \frac{c}{2}t^2 \right\}$$

$$s.t. \qquad t \geq 0$$

The Lagrangian function is $\mathcal{L} = \frac{2t(\alpha+\beta a)+a\delta t^2}{4} - \frac{c}{2}t^2 + \lambda t$. The FOC is

$$\frac{\alpha + a\beta + at\delta}{2} - ct + \lambda = 0$$

Then, the possibilities are: (i) $\lambda = 0$ and $t = \frac{\alpha+a\beta}{2c-a\delta}$, which is a candidate when $\frac{\alpha+a\beta}{2c-a\delta} \geq 0$ or, equivalently, when $c > \frac{a\delta}{2}$. (ii) $\lambda > 0$ and $t = 0$, which asks for $\frac{\alpha+a\beta}{2} < 0$, which is never the case.

Proof of Lemma 2..3 For the case $\beta = 0$ and $\delta = 1$, we consider the candidates that satisfy $a^2 - (gT + A)a + \alpha g \geq 0$ and $a^2 - (gT + A)a + \alpha g < 0$ sequentially and then we provide the solution as function of the parameters.

Step 1: $a^2 - (gT + A)a + \alpha g \geq 0$

The utility function to be considered is $u^*(t) = \alpha\left(T - \frac{a-A}{g}\right)$.

Depending on the parameters, the values of a such that $a^2 - (gT + A)a + \alpha g = 0$ (they only exist if $(gT + A)^2 - 4\alpha g \geq 0$) will lie, or not, in the interval where the junior's ability is defined, $[A, gT + A]$. When $a = A$, $a^2 - (gT + A)a + \alpha g$ has a positive value if $AT \leq \alpha$, and a negative value otherwise. When $a = gT + A$, the value for $a^2 - (gT + A)a + \alpha g$ is always positive. Hence we can define 3 possible regions to attain a solution: a) when $AT \leq \alpha$ and $(gT + A)^2 - 4\alpha g \geq 0$; b) when $AT \leq \alpha$ and $(gT + A)^2 - 4\alpha g < 0$; and c) when $AT > \alpha$ and $(gT + A)^2 - 4\alpha g \geq 0$. The One last case could be $AT > \alpha$ and $(gT + A)^2 - 4\alpha g < 0$, however it is not possible since this would mean that the function never has negative values and, simultaneously, has a negative value when $a = A$.

Case a) $AT \leq \alpha$ and $(gT + A)^2 - 4\alpha g \geq 0$

In this case the optimal ability lies in either one of the two following regions: $A \leq a \leq \frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2}$ or $\frac{gT+A+\sqrt[2]{(gT+A)^2-4\alpha g}}{2} \leq a \leq gT + A$. We first formalize

the problem with respect to the first region:

$$\begin{aligned} & \underset{a}{Max} \left\{ \alpha \left(T - \frac{a-A}{g} \right) \right\} \\ s.t \qquad a \;&\geq\; A \\ a \;&\leq\; \frac{gT + A - \sqrt{(gT+A)^2 - 4\alpha g}}{2} \end{aligned}$$

The lagrangian is $\mathcal{L} = \alpha \left(T - \frac{a-A}{g} \right) + \lambda(a - A) + \mu(\frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2} - a)$. The FOC is $-\frac{\alpha}{g} + \lambda - \mu = 0$

There are two possible cases for the lagrangian multipliers:

1) $\lambda > 0$, $\mu = 0$. $a = A$ is a candidate.

2) $\lambda > 0$, $\mu > 0$. This holds when $A = \frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2} \Leftrightarrow AT = \alpha$ which is a particular case of 1). Hence $a = A$ is again a candidate.

Formalizing the problem with respect to the second region, the FOC is:

$$-\frac{\alpha}{g} - \lambda + \mu = 0$$

One possible case exists for the lagrange multiplier:

1) $\lambda = 0$, $\mu > 0$. In this case, $a = \frac{gT+A+\sqrt{(gT+A)^2-4\alpha g}}{2}$ is a candidate.

Since the utility function is decreasing in a, $a = A$ is a candidate for the optimal ability.

Case b) $AT \leq \alpha$ and $(gT + A)^2 - 4\alpha g < 0$

The region to work with is $A \leq a \leq gT + A$, since the function always has positive values in this case. Since the utility function of the senior is decreasing in a, $a = A$ is a candidate for the optimal ability.

Case c) $AT > \alpha$ and $(gT + A)^2 - 4\alpha g \geq 0$

This is a particular case of case 1.a), where we only consider the second region, hence $a = \frac{gT+A+\sqrt{(gT+A)^2-4\alpha g}}{2}$ is a candidate for the optimal ability.

We now summarize the candidates for step 1:

$$a = A \quad \text{if} \quad AT \leq \alpha$$
$$a = \frac{gT + A + \sqrt{(gT+A)^2 - 4\alpha g}}{2} \quad \text{if} \quad AT > \alpha \text{ and } (gT+A)^2 - 4\alpha g \geq 0$$

Step 2: $a^2 - (gT + A)a + \alpha g \leq 0$

The utility function to be considered in this case is $u^*(a,t) = \frac{(at+\alpha)^2}{4a}$.

Following the same logic as in step 1, we have 2 subcases to solve this problem: 2.a) when $AT \geq \alpha$ and $(gT+A)^2-4\alpha g \geq 0$; 2.b) when $AT \leq \alpha$ and $(gT+A)^2-4\alpha g \geq 0$. The other two subcases are impossible, since $(gT + A)^2 - 4\alpha g < 0$ means the function always has positive values. Hence, in this step we take as given that $(gT+A)^2-4\alpha g \geq 0$.

Case a) $AT \geq \alpha$

There is one region for a to work with: $A \leq a \leq \frac{gT+A+\sqrt[2]{(gT+A)^2-4\alpha g}}{2}$. The maximization problem is:

$$\underset{a,t}{Max} \left\{ \frac{1}{4a} \left(a(T - \frac{a-A}{g}) + \alpha \right)^2 \right\}$$
$$s.t. \quad a \geq A \quad \text{and} \quad a \leq \frac{gT + A + \sqrt{(gT+A)^2 - 4\alpha g}}{2}$$

The FOC is:

$$\frac{1}{a^2} \left(a(T - \frac{a-A}{g}) + \alpha \right) \left(a(T - \frac{3a-A}{g}) - \alpha \right) + \lambda - \mu = 0$$

There are 4 possible cases for the lagrange multipliers:

1) $\lambda > 0$, $\mu = 0$. Looking at the FOC, it must be that $a(T - \frac{3a-A}{g}) - \alpha < 0$, that is, $a \in (\frac{gT+A-\sqrt{(gT+A)^2-12\alpha g}}{6}, \frac{gT+A+\sqrt{(gT+A)^2-12\alpha g}}{6})$, which is the case when $a = A$.

2) $\lambda = 0$, $\mu > 0$. In this case, $a = \frac{gT+A+\sqrt[2]{(gT+A)^2-4\alpha g}}{2}$ is a candidate if $a \in (\frac{gT+A-\sqrt{(gT+A)^2-12\alpha g}}{6}$, $\frac{gT+A+\sqrt[2]{(gT+A)^2-12\alpha g}}{6})$. We check that indeed a belongs to this interval. This holds only if $(gT+A)^2 - 12\alpha g \geq 0$.

3) $\lambda = 0$, $\mu = 0$. $a = \frac{gT+A\pm\sqrt[2]{(gT+A)^2-12\alpha g}}{6}$ are candidates, provided $(gT+A)^2 - 12\alpha g \geq 0$.

4) $\lambda > 0$, $\mu > 0$. This holds when $A = \frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2} \Leftrightarrow AT = \alpha$. Hence $a = A$ is a candidate again.

The utility function is decreasing from A until $\frac{gT+A-\sqrt{(gT+A)^2-12\alpha g}}{6}$, increasing from then on until $\frac{gT+A+\sqrt{(gT+A)^2-12\alpha g}}{6}$, and decreasing onwards. Hence, when $(gT + A)^2 - 12\alpha g \geq 0$ $a = \frac{gT+A+\sqrt{(gT+A)^2-12\alpha g}}{6}$ is candidate for the optimal ability and when $(gT+A)^2 - 12\alpha g < 0$, $a = A$.

Case b) $AT \leq \alpha$

There is one region for a to work with: $\frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2} \leq a \leq \frac{gT+A+\sqrt{(gT+A)^2-4\alpha g}}{2}$.

The FOC is the same as in case a), hence the possible cases for the lagrangean multipliers are:

1) $\lambda > 0, \mu = 0$. $a = \frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2}$ and it must be that $a \in (\frac{gT+A-\sqrt{(gT+A)^2-12\alpha g}}{6}$, $\frac{gT+A+\sqrt{(gT+A)^2-12\alpha g}}{6})$. Since $\frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2} < \frac{gT+A-\sqrt{(gT+A)^2-12\alpha g}}{6}$, indeed it is a candidate.

2) $\lambda = 0, \mu > 0$. $a = \frac{gT+A+\sqrt{(gT+A)^2-4\alpha g}}{2}$ and it must be that $a \in (\frac{gT+A-\sqrt{(gT+A)^2-12\alpha g}}{6}$, $\frac{gT+A+\sqrt{(gT+A)^2-12\alpha g}}{6})$. It is straightforward to check that a indeed belongs to this interval, so it is a candidate, provided that $(gT+A)^2 - 12\alpha g \geq 0$.

3) $\lambda = 0, \mu = 0$. In this case, $a = \frac{gT+A \pm \sqrt{(gT+A)^2 - 12\alpha g}}{6}$ provided that $(gT + A)^2 - 12\alpha g \geq 0$.

4) $\lambda > 0$,$\mu > 0$. In this case, $a = \frac{gT+A}{6}$, which happens when $(gT+A)^2 - 12\alpha g = 0$, a particular case of 3).

Analyzing all the candidates and the behavior of the utility function, $a = \frac{gT+A+\sqrt{(gT+A)^2-12\alpha g}}{6}$ is a candidate for the optimal ability when $(gT + A)^2 - 12\alpha g \geq 0$ and $a = \frac{gT+A-\sqrt{(gT+A)^2-4\alpha g}}{2}$ when $(gT + A)^2 - 12\alpha g < 0$.

We now summarize all candidates for case 2:

$$a = \frac{gT + A + \sqrt{(gT + A)^2 - 12\alpha g}}{6} \quad \text{if} \quad (gT + A)^2 - 12\alpha g \geq 0$$

$$a = A \quad \text{if} \quad AT \geq \alpha \text{ and } (gT + A)^2 - 12\alpha g < 0$$

$$a = \frac{gT + A - \sqrt{(gT + A)^2 - 4\alpha g}}{2} \quad \text{if} \quad AT < \alpha \text{ and } (gT + A)^2 - 12\alpha g < 0$$

As function of the parameters, we evaluate and compare the utility of the solutions attained in step 1 and step 2. The final solutions for a^* and t^* (attained recursively) are:

$(gT + A)^2 - 12\alpha g \geq 0$	$a^* = \frac{gT+A+\sqrt{(gT+A)^2-12\alpha g}}{6}$	$t^* = \frac{5(gT+A)-\sqrt{(gT+A)^2-12\alpha g}}{6g}$
$(gT + A)^2 - 12\alpha g < 0$	$a^* = A$	$t^* = T$

REFERENCES CITED

1. Armbruster, L., 2008, "The Rise of The Post-Doc as Principal Investigator? How PHDs May Advance in Their Career and Knowledge Claims in the New Europe of Knowledge", *Policy Futures in Education* 6(4): 409-423.

2. Banal-Estañol, A. and Macho-Stadler, I., 2010, "Scientific and Commercial Incentives in R&D: Research vs. Development", *Journal of Economics and Management Strategy* 19(1): 185-221.

3. Cech, T. and Bond, E., 2004, "Managing Your Own Lab", *Science* June 18, Vol. 304: 1717.

4. European Commission, 2004, "*Increasing Human Resources for Science and Technology in Europe*", Report chaired by J.M. Gago, High Level Group (HLG) Human Resources for Science and Technology in Europe.

5. Frame, I. and Allen, L., 2002, "A Flexible Approach to PhD Research Training", *Quality Assurance in Education* 12(2): 98-103.

6. Golde, C., 2000, "Should I stay or should I go? Student descriptions of the doctoral attrition process", *The Review of Higher Education* 23(2): 199-227.

7. Green, S. and Bauer, T., 1995, "Supervisory Mentoring by Advisers: Relationships With Doctoral Student Potential, Productivity and Commitment", *Personnel Psychology* 48(3): 537-561.

8. Kram, K., 1983, "Phases of the Mentor Relationship", *Academy of Management Journal* 26: 608-625.

9. Kram, K. and Isabella, L., 1985, "Mentoring Alternatives: The Role of Peer Relationships in Career Development", *Academy of Management Journal* 28(1): 110-132.

10. Holmstrom, B. and Milgrom, P., 1991, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design", *Journal of Law, Economics and Organization* 7: 24-51.

11. Levinson, D., 1978, "*The season's of a man's life*", New York: Knopf.

12. Lacetera, N. and Zirulia, L., 2008, "Knowledge Spillovers, Competition, and Taste for Science in a Model of R&D Incentive Provision", Universita' di Bologna, Working Paper.

13. Lovitts, B. E., 2001, "*Leaving the ivory tower: the causes and consequences of departure from doctoral study*", Lanham, MA, Rowman & Littlefield.

14. Lovitts, B. E., 2005, "Being a good course-taker is not enough: a Theoretical Perspective on the Transition to Independent Research", *Studies in Higher Education* 30(2): 137-154.

15. Nerad, M. and Cerny, J., 1999, "Postdoctoral Patterns, Career Advancement, and Problems", *Science* September 3, Vol. 285: 1533-1535.

16. Paglis, L., Green, S. and Bauer T., 2006, "Does Advisor Mentoring Add Value? A Longitudinal Study of Mentoring and Doctoral Student Outcomes", *Research in Higher Education* 47(5): 451-476.

17. Pole, C., Sprokkereef, A., Burgess, Robert G., and Lakin, E., 1997, "Supervision of Doctoral Students in the Natural Sciences: expectations and experiences", *Assessment & Evaluation in Higher Education* 22(1): 49-63.

18. Puljak, L., 2006, "Career Blocker: Bad Advisors" in http://sciencecareers.sciencemag.org.

19. Stephan, P. E. and Levin,S.G., 1992, Striking the Mother Lode in Science: The Importance of Age, Place, and Time, New York: Oxford University Press.

20. Vogel, G., 1999, "A Day in the Life of a Topflight Lab", *Science* September 3, Vol. 285: 1531-1532.

21. Walckiers, A., 2008, "Multidimensional Contracts with Task-Specific Productivities: An Application to Universities", *International Tax and Public Finance* 15: 165-198.

CHAPTER 3.

PATENT STRATEGY OF PHARMACEUTICALS: WHEN PAY-FOR-DELAY SETTLEMENTS DELAY NEW DRUGS

3.1 Introduction

This article analyzes the effects of *pay-for-delay* settlements on the patent strategy of brand pharmaceuticals when facing potential competition from generic firms before brand drug patent expiration, and how it is shaping the innovation in the pharma industry.

Competition of generic firms in brand drug markets before patent expiration has been made possible in the U.S. in 1984 with the Hatch-Waxman Act. The Act introduced a new regulation to the Food & Drug Administration (FDA) that launched a generic approval pathway called Abbreviated New Drug Application (ANDA). By filing an ANDA, generic firms don't have to undergo expensive clinical trials to demonstrate the generic drug's bioequivalence to the brand drug to obtain FDA approval. This introduced a major change in the industry, since generics don't incur in massive costs to enter the market. Generics can apply an ANDA and enter in a brand drug market before patent expiration through the paragraph IV certification of the Act. Paragraph IV states that a generic drug can enter before patent expiration by proving that the patent of originator brand drug is not infringed or that is invalid. Doing so, to the first generic to file an ANDA is given an exclusive right to 180-day marketing period, which implies high profits during this period.[1] Upon entry notification of the first-filer, the

[1] Revenues for the first-filer are very high during this period when compared to the period after, since competition by other generics drives down the drug price.

brand firm is entitled, within a period of 45 days, to either file an infringement suit or accommodate entry. Infringements suits either end up in a trial case or in a settlement agreement.[2]

Settlements usually take the form of license agreements where the licensor receives a fee, in the form of a fixed fee or a royalty, for transfering the technology to the licensee. However in recent years a new kind of settlements has emerged in the pharmaceutical industry to resolve patent disputes between brand firms and generic firms: the *pay-for-delay* settlements. *Pay-for-delay* settlements are settlements where a brand firm pays a fixed sum of money to the generic firm in order to stay out of the market, and the parties agree to dismiss the patent lawsuit. The agreement may also contain a clause where the generic enters before the expiration date of the brand drug patent (Hemphill, 2006).

There has been a significant increase in the number of this type of agreements in the U.S. since 2003, the year that the Federal Trade Comission (FTC) started to officially track them. In an FTC report of the fiscal year 2011, which ran from October 1, 2010 through September 30, 2011, the FTC reported that 54 patent settlements were made between brand firms and generic firms, out of which 28 were potentially *pay-to-delay* settlements, that is, they contained a compensation payment and a restriction on entry.[3] Out of the 28 deals, 18 involved generics that were first-filers, which means

[2]Empirical studies show that while the exposure to litigation is low on average, arround 1%, the pharmaceutical sector shows much higher risk, especially for "valuable" drugs and health patents, where the estimated probability of litigation during the lifetime of the patent rises to more than 25% (Lanjouw and Schankerman, 2001 and 2004). Most of the infringement suits, arround 95%, end up in a settlement instead of a trial sentence (Lanjouw and Schankerman, 2001 and 2004).

[3]This number is similar to the 31 settlements of the same type in the fiscal year 2010.

they were first to seek FDA approval to market the generic drug and to be eligible for market exclusivity. The 28 deals involve drugs with $9 billion of combined sales in 2010, approximately 14% of U.S. combined sales of the 200 top drugs in that year. Hence, *pay-for-delay* settlements are associated with non-negligible markets. Moreover, out of the remaining 128 deals, 100 restrict the generic manufacturer's ability to market but do not contain any explicit compensation and 28 have no restrictions on entry to generic firms.

In the U.S., the debate over the use of pay-for-delay settlements to resolve paragraph IV patent disputes is very strong.[4] It is centered on FTC antitrust allegations that these settlements eliminate potential competition of brand firms and do not allow access of lower priced drugs to consumers by delaying the entry time of generics (FTC Statement, 2009). Antitrust enforcement actions have taken place in recent years, but courts keep a lenient approach and these agreements have become common industry strategy. The FTC has continued its allegations and supported a Bill on a complete ban of *pay-for-delay* agreements in 2011. This Bill did not pass but still the FTC is supporting the preparation of a similar bill to be discussed in the U.S. congress in 2012.

On the other side of this debate are the brand pharmaceuticals. The Pharmaceutical Research and Manufacturers of America (PHRMA) refers that "*a complete ban on pay for delay settlements could decrease the value of patents and reduce incentives for future innovation of new medicines*" (PhRMA statement, 2010). Also, the Generic Pharmaceutical Association (GhRMA) is in favor of these settlements, as in their point

[4] In Europe the concern and debate over this type of settlement is still building up since the European Commission has started to highlight these deals in its 2009 year-long pharma inquiry.

of view these settlements allow the introduction of generics earlier than would otherwise have been possible.

This motivation of article is builds upon the brand pharmaceutical association's argument. Focusing on the incentives to invest in new drug discovery, the research question is: how are *pay-for-delay* settlements affecting pharmaceuticals' patenting decisions? I offer a new insight on the role that *pay-for-delay* settlements are possibly already playing in pharmaceuticals' patent strategy and help explain the decreasing trend of new drugs being lauched to the market.

To perform the analysis, I model a brand pharmaceutical that decides either to patent and market a new drug or to improve protection of an existing drug (invest in patents that contain claims to support the initial patent, new therapeutical treatments, new dosages, among others). I analyse the patent decision under two different settings: when *pay-for-delay* settlements are allowed vs when *pay-for-delay* settlements are not allowed. If not allowed, I consider that firms can perform *regular* settlements instead, i.e., technology licensing agreement to the generic firm.

This article shows that, in some cases, *pay-for-delay* settlements increase generic entry due to higher prospective rents that a generic first-filer receives in order not to compete. Also, entry of generics is more frequent when patent protection is low and intermediate. More importantly, this entry induces brand pharmaceuticals to preemptively shift their patent decisions, by favoring protection of the existing drug and in detriment of new drug discovery. I predict that this result is more frequent for larger markets or when the generic firm faces a sufficiently low entry cost. This results in the delay of new drug development and may help explain the significant decrease in

new drug launches in recent years. Hence, a ban on *pay-for-delay* settlements can help mitigate the delay in new drug developments.

The model is in the spirit of Crampes and Langinier (2002) when I consider the reaction of an incumbent firm facing entry. However, to simplify the analysis I assume firms hold symmetric information with respect to the outcome of litigation, and I do not consider an inefficient judicial system. Instead, the model's approach is the protection level of patents. The novelty in this paper is to model the patent strategy of an firm under the threat of future entry and dispute. This paper draws conclusions on the effect that different types of settlement agreements have on the brand drug firm's decision of what to patent, specifically between *regular* settlements and *pay-for-delay* settlements.

My analysis fits in the economic literature on the antitrust issues of *pay-for-delay* settlements in the U.S. pharmaceutical industry. When arguing that such settlements are anti-competitive, Shapiro (2003) relies on the proposition that a settlement should not lead to lower expected consumer surplus than would arise from ongoing litigation. Willig and Bigelow (2004) argue that such reverse payments can be pro-competitive, in the presence of risk aversion, imperfect capital markets and asymmetric information about the economic life of the patent. However, as Schrag (2007) argues, such settlements can harm consumers when further entry is considered, since they undermine subsequent entrants' incentive to challenge the patent. My approach is to take into account that new markets may not be available to consumers due to the protection effect that *pay-for-delay* settlements cause.

Also, the paper is related to the literature on patent litigation and strategic patenting. The literature on patent litigation was pioneered by Meurer (1989) by

adressing patent licensing induced to avoid litigation. In more recent literature, the usual argument for firms' incentives to settle is the bargaining surplus created by a settlement, as it avoids the large costs litigation. Several authors have studied how the possibility of future litigation can affect entry decisions and settlements when the outcome of a trial is uncertain (Aoki and Hu, 1999, and Crampes and Langinier, 2002). Also, in a paper probably closest to this work, Bessen and Meurer (2006) have modelled the incentives to invest in protection of a patent under the possibility of future entry and litigation.

In the strategic patenting literature, studies have focused on two main reasons to patent: blocking reasons and exchange reasons. Kash and Kingston (2001) identify that simple technologies, such as the pharmaceutical, use patents for their traditional purpose of preventing other firms from using the invention. Cohen (2002) explore U.S. and Japan data on two reasons for patenting: building a patent fence and patenting to trade. Their results also show substantial differences in patent strategies by technology complexity . Blind and Köhler (2010) survey approximately 440 german firms on the effect of patenting motives on patent portfolio characteristics and find evidence of amendments for exhange reasons, but not for protection reasons.

The rest of the paper is organized as follows. I introduce the model in Section 3.2. In Section 3.3, as a benchmark, I present the equilibrium when *pay-for-delay* settlements are not allowed. The main analysis where *pay-for-delay* settlements are allowed can be found in Section 3.4. The impact of allowing *pay-for-delay* settlements is presented in Section 3.5 and Proposition 3..1 states the main result, as well as consumer welfare analysis. To illustrate the results, I offer a simple Cournot competition example in

Section 3.6. In Section 3.7 several discussions are provided. I conclude in Section 3.8. All proofs are in the Appendix.

3.2 The model

A brand pharmaceutical firm, O, is the owner of patent A_0 that protects drug A for which O is the market incumbent. I denote this market as A. O faces a potential entrant in this market, a generic firm G, under Paragraph IV of the Hatch-Waxman Act. O decides, at stage 1, either to invest in protecting more drug A, by developing a patent A_1, or invest in developing a patent B to protect a new molecular entity[5] that allows to sell a new drug in a market also denoted as B. If O develops patent B, it holds a portfolio composed of patents A_0 and B, denoted as $A_0 + B$.

To develop patent B and market the new drug, O must spend R_B in R&D. The investment is assumed to be profitable, that is, the earnings attained with this new drug always exceeds the development cost. In this case, O holds patent portfolio $A_0 + B$ and drug A is protected only by patent A_0, that has strength α_0 and represents the probability that O wins a patent trial case against an infringing firm in market for drug A, $\alpha_0 \in [0, 1]$. Patent A_1, on the other hand, improves the protection of brand drug A. To develop A_1 O must spend R_A in R&D. In the model I assume developing A_1 provides only protection benefits in disputes to firm O, i.e., O's profits do not change. Hence, patent A_1 has strength $\alpha_1 \in (\alpha_0, 1]$.

Generic firm G decides whether to enter market A at the beginning of stage 2

[5] A new molecular entity is, according to the U.S. Food and Drug Administration (FDA), a drug that contains no active ingredient that has been approved by the FDA in any other application submitted under the U.S. Federal Food, Drug, and Cosmetic Act.

by filing an ANDA application. Entry implies a sunk cost $F > 0$. One may wonder whether G could also enter market for drug B if O decided to develop it. However the Hatch-Waxman Act protects patents on New Molecular Entities by giving 5-year challenge immunity from generic drugs, so I exclude this possibility in the model. I assume players hold symmetric information, i.e., there is common knowledge about the protection level of drug A.

If G enters market A, O can either accommodate or sue for infringement at stage 3. In case O accommodates entry, firms share market A in duopoly. If the parties go to trial, each faces a trial cost $K \geq 0$. If O wins the case, the patent is upheld and G must exit market A. If, on the other hand, O loses the case then it shares market A in duopoly with G.

If the outcome is settlement, O incurs in settlement cost $C \geq 0$ by handling all legal procedures concerning the licensing contract (as in Crampes and Langinier, 2002). If the settlement is a *regular* licensing agreement, then O receives a fixed license fee L_α from G in return for sharing the market in duopoly, $\alpha = \alpha_0, \alpha_1$. If the settlement takes the form of *pay-for-delay*, O pays G a fixed license fee L_α^{PD} in return for G not to compete in market A, $\alpha = \alpha_0, \alpha_1$. The license fee, in either type of settlement that firms might engage, is negotiated through Nash Bargaining, with exogenous bargaining power $\rho > 0$. [6] Without loss of generality, I assume firms have the same bargaining power, i.e., $\rho = 0.5$.

I denote the monopoly and duopoly profits in market for drug A as π^M and π^D,

[6] Although *pay-for-delay* contracts in practice may include a time delay of entry of the generic drug firm, I do not assume any explicit time delay. The bargaining power can also be interpreted as a reduced form of the time delay negotiation with the entrant.

respectively, with $\pi^M > 2\pi^D$. I assume that the product market for drug A can always accommodate two firms, i.e., $\pi^D - F \geq 0$.[7] In the market for drug B, I denote the monopoly profit as π^B. Finally, I denote the payoffs under accommodation, settlement and trial as Ac_j, S_j and T_j, respectively, for firm $j = O, G$. The model has three stages.

- In stage 1, O decides on its patent strategy. O either invests in more protection of drug A by developing patent A_1, or invests in a new drug by developing patent B.

- In stage 2, G decides whether or not to enter market A. If G does not enter the game ends. If G enters, the game proceeds to stage 3.

- In stage 3, after G enters market A, O decides either to accommodate entry or sue G. If O sues G, they either settle or go to trial.

We solve the model by backward induction and the equilibrium concept is the Subgame Perfect Nash Equilibrium.

3.3 Benchmark: *pay-for-delay* settlements are not permitted

The objective of this section is to evaluate O's patent strategy when pay-for-delay settlements are not legally permitted. In this setting, O faces possible entry of G in market A. If O decides to sue, there are two possible outcomes, a *regular* settlement or trial. I name *regular* to the settlement where the entrant pays a license fee to the

[7] This way, I rule out of the analysis the cases where generic firms do not enter the market in a "pacific" way (without dispute from incumbents) either because the market is too small or because development costs of generic version of the brand drug are too prohibitive.

incumbent in order to access the technology and compete in the market.[8]

Stage 3 This stage only exists if G enters market A at stage 2. For both firms, the incentive to settle is the bargaining surplus it creates. As O's payoffs only change due to different protection levels, the thresholds for which O chooses to accommodate, settle or go to trial are the same for both strategies. So, I solve for a general protection level α, for $\alpha = \alpha_0, \alpha_1$.

When O's profit from accomodating is at least as the profit from going to trial, i.e., $K \geq \alpha(\pi^M - \pi^D)$, then the reservation value for any settlement is the duopoly profit, since if negotiations were to fail, O would choose accommodation. Since there is no risk of trial, no bargaining surplus exists and settlement is a not mutually profitable alternative, so O accommodates. On the other hand, if $K < \alpha(\pi^M - \pi^D)$ then trial is a "back-up solution" to settlement. Under settlement, the optimal license fee is $L^*_\alpha = 0.5\alpha\pi^M$, $\alpha = \alpha_0, \alpha_1$. So, if O develops A_1 and settles, it earns $S_O = \pi^D + L^*_{\alpha_1} - R_A - C$. If O develops B, the payoff is $S_O = \pi^D + L^*_{\alpha_0} + \pi^B - R_B - C$. The decision to settle or go to trial is made by comparing the payoffs. The settlement and trial thresholds expressed in terms of patent strength are denoted as $\alpha_S \equiv \frac{K}{\pi^M - \pi^D}$ and $\alpha_T \equiv \frac{2(K-C)}{\pi^M - 2\pi^D}$. Lemma 3..1 describes the equilibrium for a general protection level.

Lemma 3..1 *When pay-for-delay settlements are not allowed, if G has entered at stage 2, then, at stage 3, given $(\alpha, entry)$ for strength $\alpha = \alpha_0, \alpha_1$, O accommodates entry for $\alpha \in (0, \alpha_S]$, settles for $\alpha \in (\alpha_S, \alpha_T]$ and goes to trial for $\alpha \in (\alpha_T, 1]$.*

[8] The strength of patent A_1, α_1, and of portfolio $A_0 + B$, α_0, dictates whether O accomodates or sues. One may also consider the benchmark case where no kind of settlements are permitted. However settlements are known to be pareto improving for high enough litigation costs. One could mimic this case with low enough litigation costs, however I consider this case less realistic.

Figure 3..1 is a graphical description of O's reaction to entry in market A. Holding the level of protection fixed, O's willingness to dispute entry increases with the profitability of drug A. If A is a *blockbuster* drug (either with high profit margin or a large market), trial also becomes more likely. However, higher litigation costs mitigate O's desire to defend drug A in trial, choosing rather to accommodate more and settle more. Also, the relative desire to settle over going to trial is inversely affected by settlement costs.

Figure 3..1: Protection level $\alpha \in \{\alpha_0, \alpha_1\}$

The result in Lemma 1 can also be expressed in terms of costs of dispute. O has the incentive to accommodate entry if $K \geq \alpha(\pi^M - \pi^D)$, otherwise to sue. In the dispute region, O settles if the costs are not relatively high with respect to trial costs, $C \leq K - 0.5\alpha(\pi^M - 2\pi^D)$, otherwise goes to trial.

Stage 2 G's decision as to whether enter of not in stage 2, as a funcion of $\alpha \in \{\alpha_0, \alpha_1\}$, is the following. When patent strength is $\alpha \in (0, \alpha_S]$, G always enters in the market for drug A. When $\alpha \in (\alpha_S, \alpha_T]$, G enters as long as its profits are positive, so it enters for protection levels below the entry threshold θ_S, where $\theta_S \equiv \frac{2(\pi^D - F)}{\pi^M}$. Entry is conditional on $\theta_S > \alpha_S$, i.e., on a low enough entry cost $F < F_S$, where F_S is the entry cost threshold. When $\alpha \in (\alpha_T, 1]$, G enters for a positive expected profit, i.e., for patent strength below the threshold $\theta_T \equiv 1 - \frac{K+F}{\pi^D}$. Entry is also conditional on

$\theta_T > \alpha_T$, i.e., on $F < F_T$, where F_T is the entry cost threshold. The following lemma summarizes the entry decision. From direct calculation one can observe that $F_T < F_S$ always holds.

Lemma 3..2 *Given $\alpha \in \{\alpha_0, \alpha_1\}$, at stage 2:*

When $\alpha \in [0, \alpha_S]$, G enters market A. When $\alpha \in (\alpha_S, \alpha_T]$, G enters for $\alpha \leq \theta_S$ and $F < F_S$, and does not otherwise. When $\alpha \in (\alpha_T, 1]$, G enters for $\alpha \leq \theta_T$ and $F < F_T$, and does not otherwise.

Lemma 3..2 states that the entry region increases as the entry cost faced by G in market for brand drug A decreases. When entry cost is high enough, $F > F_S$, G only enters market A if protection α is low enough and O always accommodates the ANDA approval of the generic, i.e., when $\alpha \in [0, \alpha_S]$. This might be related to high development costs of the generic version of drug A that only allows G to enter only if no dispute exists, otherwise would imply extra dispute costs. Also, keeping the development costs fixed, it may be because market A is not big enough to induce generic firms to enter and face dispute.If $F < F_T$, G is willing to go to trial in order to compete in market A. Overall, all things being equal, profitability of drug A increases the chances of generic competition, being *blockbuster* drugs the most proeminent candidates to face generic competition.

Figure 3..2 provides a graphical interpretation of the result in Lemma 2 when $F_T < F < F_S$. Since $\alpha_1 > \alpha_0$, we consider the triangle below the 45^o line. There are, at most, six equilibrium regions. The full lines separate them. When $\alpha \in [0, \alpha_S], \alpha =$

α_0, α_1, we have the region where O accommodates in any case, $AcAc$ (accommodate/accommodate). Where α_1 and α_0 are such that O settles and accommodates, respectively, the region is denoted as SAc. When α_1 is such that O goes to trial and α_0 such that O accommodates, is region TAc. When both α_0 and α_1 are such that O settles is region SS. Also, when α_1 is such that O goes to trial and α_0 such that O settles is the region TS. Finally, when both α_0 and α_1 are such that O goes to trial, the region is TT. Also, we denote the region of entry as E and the region of no entry as NE. As $F_T < F < F_S$, G enters for protection levels lower than θ_S, represented by the dotted lines, but not for higher patent protection. Holding development costs fixed, it must be the case that market A is attractive enough to justify entering and negotiating with O.

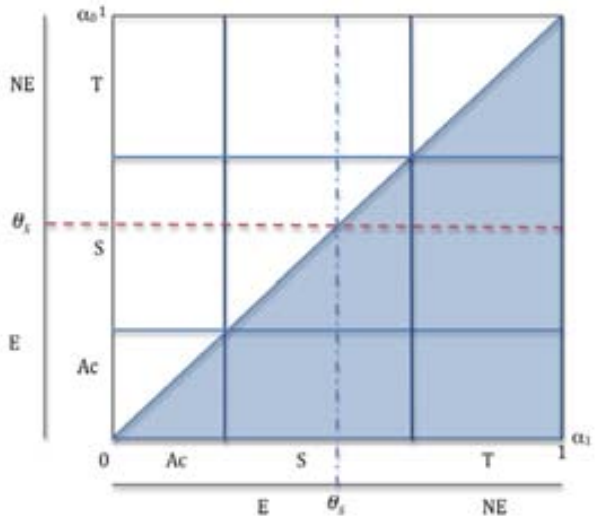

Figure 3..2: G's entry when $F_T < F < F_S$

Stage 1 O's patenting decision between A_1 and B is made foreseeing protection levels α_0 and α_1 and G's entry decision, which depends on F. Lemma 3..3 describes the

equilibrium. In each region, when G files an ANDA approval, O's decision of what innovation to follow is based on protection level thresholds, that I denote by $\Psi \equiv \frac{2(\pi^B - R_B + R_A)}{\pi^M}, \psi \equiv \frac{2(\pi^M - \pi^D - R_A - \pi^B + R_B + C)}{\pi^M}, \gamma \equiv \frac{2(\pi^B - R_B + R_A + C)}{\pi^M}, \varpi \equiv \frac{\pi^B - R_B + R_A}{\pi^M - \pi^D}, \Gamma \equiv \frac{2(\pi^B - R_B + R_A + K - C)}{\pi^M}$ and $\lambda \equiv 2(1 - \frac{\pi^D}{\pi^M})$. It is intuitive that these thresholds favour the choice of protecting more brand drug A when the market at stake is attractive.

Lemma 3..3 *Given $((\alpha_0, \alpha_1), F)$, the solution of the game is:*

i) *In region AcAc, O chooses B.*

ii) *If $F > F_S$, B is chosen: in regions SAc and TAc if $\pi^B - R_B \geq \pi^M - \pi^D - R_A$; in SS, TT and TS. Otherwise, A_1 is chosen.*

iii) *If $F_T < F < F_S$, B is chosen: in region SS if $\alpha_0 \in [\alpha_1 - \Psi, \theta_S]$ and $\alpha_1 \leq \theta_S$ or if $\alpha_0 \in [\psi, \theta_S]$ and $\alpha_1 \geq \theta_S$; in SAc if $\alpha_1 \leq Min\{\theta_S, \gamma\}$; in SAc and TAc if $\alpha_1 \geq \theta_S$ and $\pi^B - R_B \geq \pi^M - \pi^D - R_A$; in TS if $\alpha_0 \in [\psi, \theta_S]$ and $\alpha_1 \geq \theta_S$; and in SS, TT and TS, if $\alpha \geq \theta_S$, $\alpha = \alpha_0, \alpha_1$. If not, A_1 is chosen.*

iv) *If $F < F_T$, B is chosen: in region SAc if $\alpha_1 \leq \gamma$; in SS if $\alpha_1 - \alpha_0 \leq \Psi$; In TAc if $\alpha_1 \leq Min\{\theta_T, \alpha_S + \varpi\}$ or if $\alpha_1 > \theta_T$ and $\pi^B - R_B \geq \pi^M - \pi^D - R_A$; in TS if $\alpha_1 \leq Min\{\theta_T, \alpha_0 + \Gamma\}$ or if $\alpha_1 > \theta_T$ and $\alpha_0 \geq \psi$; in TT, if $\alpha_1 \leq Min\{\theta_T, \alpha_0 + \varpi\}$, if $\alpha_0 \in [1 + \alpha_S - \varpi, \theta_T]$ and $\alpha_1 > \theta_T$, or if $\alpha \geq \theta_T$, $\alpha = \alpha_0, \alpha_1$. If not, A_1 is chosen.*

When there is no entry under dispute, O chooses to develop B since more protection of drug A is valueless. Also, if initial patent strength α_0 is already significant, additional patent protection brings little value and drug B is relatively more attractive.

Higher profitability of drug A, ceteris paribus, induces more generic entry and more protection when improving drug A $(\alpha_1 - \alpha_0)$ is significant enough.

3.4 *Pay-for-delay* settlements are permitted

Now *pay-for-delay* settlements are allowed as an outcome of dispute between O and G. In this setting, O is allowed to settle by paying a licensing fee to G in order to keep G from competing in the market for drug A.

Stage 3 I solve for α, $\alpha = \alpha_0, \alpha_1$. This stage only exists if G enters. If $K \geq \alpha(\pi^M - \pi^D)$, O accommodates. On the other hand, if $K < \alpha(\pi^M - \pi^D)$, O is willing to dispute entry. Under settlement, the optimal license fee is $L_\alpha^{PD*} = 0.5(1-\alpha)\pi^M$, $\alpha = \alpha_0, \alpha_1$. If O develops A_1 and settles, it earns $S_O = \pi^M - L_{\alpha_1}^{PD*} - R_A - F$. If O develops B, the payoff is $S_O = \pi^M - L_{\alpha_0}^{PD*} + \pi^B - R_B - C$. In terms of protection, the settlement and trial thresholds are $\alpha_S^{PD} \equiv \frac{K}{\pi^M - \pi^D}$ and $\alpha_T^{PD} \equiv 1 + \frac{2(K-C)}{\pi^M - 2\pi^D}$. Since $\alpha_S^{PD} = \alpha_S$ we use α_S in both settings to avoid excessive notation. Lemma 3..4 describes the equilibrium.

Lemma 3..4 *When Pay-for-delay settlements are possible, if G has entered at stage 2, at stage 3 given (α, entry) for $\alpha = \alpha_0, \alpha_1$, O accommodates entry when $\alpha \in (0, \alpha_S]$, settles when $\alpha \in (\alpha_S, \alpha_T^{PD}]$ and goes to trial when $\alpha \in (\alpha_T^{PD}, 1]$.*

In terms of dispute costs, O accommodates entry if $K \geq \alpha(\pi^M - \pi^D)$, otherwise settles. In the dispute region, O settles if $C \leq 0.5(1-\alpha)(\pi^M - 2\pi^D) + K$, otherwise goes to trial. The level of protection for which O is indifferent between accommodation and settlement does not change with the type of settlement being negotiated. However, as it is possible to negotiate a *pay-for-delay* settlement with G, O is now indifferent

between settlement and trial at a higher protection level since it is always the case that $\alpha_T^{PD} > \alpha_T$. Moreover, it can never be the case that the 3 regions of reaction coexist in both settings, *pay-for-delay* and *regular* settlements. In other words, in the case O settle or goes to trial under *regular* settlements, O is only willing to perform *pay-for-delay* settlements in a dispute with the generic firm.

Stage 2 Given $\alpha \in \{\alpha_0, \alpha_1\}$, for $\alpha \in (0, \alpha_S]$ G always enters market A. When $\alpha \in (\alpha_S, 1]$, O settles and G enters if its profits are positive. I denote the entry threshold as $\theta_S^{PD} \equiv 1 - \frac{2F}{\pi^M}$. Entry is conditional on $\theta_S^{PD} > \alpha_S$, i.e., an entry cost $F < F_S^{PD}$ where F_S^{PD} is the entry cost threshold. When $\alpha \in (\alpha_T^{PD}, 1]$, G enters for a positive expected profit, i.e., for patent strength below $\theta_T \equiv 1 - \frac{K+F}{\pi^D}$. Entry is conditional on $\theta_T > \alpha_T^{PD}$, i.e., an entry cost $F < F_T^{PD}$. The following lemma describes the entry decision.

Lemma 3..5 *For $\alpha \in \{\alpha_0, \alpha_1\}$, at stage 2:*

When $\alpha \in [0, \alpha_S]$, G enters market for drug A. When $\alpha \in (\alpha_S, \alpha_T^{PD}]$, G enters for $\alpha \leq \theta_S^{PD}$ and $F < F_S^{PD}$, and does not otherwise. When $\alpha \in (\alpha_T^{PD}, 1]$, G enters for $\alpha \leq \theta_T$ and $F < F_T^{PD}$, and does not otherwise.

As expected, the profitability of drug A increases the chances of generic competition under *pay-for-delay* settlements as well. The particularity of this setting is that G may enter more often in market A under *pay-for delay* settlements for levels of protection where O settles, as $\theta_S^{PD} > \theta_S$ always holds (which implies that $F_S^{PD} > F_S$).

Corollary 3..1 *Firm G files an ANDA more often under pay-for-delay settlements than under regular settlements. Formally, $\theta_S^{PD} > \theta_S$.*

The main reason is that the generic firm G antecipates to extract more rents from the brand drug pharmaceutical by negotiating an attractive compensation (or an antecipated entry date before patent expiration) in exchange for not competing until then.

Stage 1 Since O's strategy is unchanged in $AcAc$, I only describe the remaining equilibrium regions. I exclude from the description the case where the entry cost is too high for G too enter $F > F_S^{PD}$, since solution coincides with lemma 3 when there is no entry under dispute. In each region, when G files an ANDA, O's decision of what innovation to follow is based on threshold levels for patent protection, denoted as $\Psi, \Gamma, \lambda, \delta, \varpi$ (described in section 3.3) and $\Phi \equiv 1 - \frac{2(\pi^B - R_B + R_A - C)}{\pi^M}$, $\Omega \equiv 1 - \frac{2(\pi^B - R_B + R_A + C + \pi^D)}{\pi^M}$.

Lemma 3..6 *Given $((\alpha_0, \alpha_1), F)$, the solution of the game is:*

i) *If $F_T^{PD} < F < F_S^{PD}$, B is chosen: in region SS if $\alpha \leq \theta_S^{PD}$ and $\alpha_1 - \alpha_0 \leq \Psi$ or if $\alpha_0 \in \left[\Phi, \theta_S^{PD}\right]$ and $\alpha_1 > \theta_S^{PD}$; in SAc if $\alpha_1 \leq Min\left\{\theta_S^{PD}, \Omega\right\}$ or $\alpha_1 > \theta_S^{PD}$ and $\pi^B - R_B \geq \pi^M - \pi^D - R_A$; In TS if $\alpha_0 \geq \alpha_1 \lambda - \delta$; and in TT. If not, A_1 is chosen.*

ii) *If $F < F_T^{PD}$, B is chosen: in SS if $\alpha_1 - \alpha_0 \leq \Psi$; in SAc if $\alpha_1 \leq \Omega$; in TT if $\alpha_0 \in \left[\alpha_1 - \varpi, \theta_T^{PD}\right]$ and $\alpha_1 \leq \theta_T^{PD}$ or if $\alpha_0 \in \left[1 + \alpha_S - \varpi, \theta_T^{PD}\right]$ and $\alpha_1 > \theta_T^{PD}$ or if $\alpha > \theta_T^{PD}$, $\alpha = \alpha_0, \alpha_1$; in TAc if $\alpha_1 \leq Min\left\{\theta_T^{PD}, \frac{\alpha_0 + \Gamma}{\lambda}\right\}$ or if $\alpha_1 > \theta_T^{PD}$ and $\pi^B - R_B \geq \pi^M - \pi^D - R_A$; and in TS if $\alpha_1 \leq Min\left\{\theta_T^{PD}, \frac{\alpha_0 + \delta}{\lambda}\right\}$ or if $\alpha_1 > \theta_T^{PD}$ and $\alpha_0 \geq \Phi$. If not, A_1 is chosen.*

Compared with *regular* settlements, the relevant threshold levels to decide between A_1 and B show that in some cases O's decision is in favor of A_1, while others favor more the new drug B. Also, there are cases where no effect is produced at all, as in region SS where O faces the same criterion in both regimes. Hence, the overall effect of *pay-for-delay* settlement agreements on patenting strategy is uncertain. However, the magnitude of differences in both regimes is higher with the profitability of brand drug A. In the following section I focus on a particular case and identify when *pay-for-delay* settlements favour protection improvement of drug A when compared to *regular* settlements.

3.5 *Pay-for-delay* settlements and patent strategy

This section shows how *pay-for-delay* settlements affect entry of competing generic drugs, as well as the patent decision of brand firms. Of course, this influences the innovation path of the entire phamaceutical industry and, hence, of the health sector.

Starting from the benchmark where *pay-for-delay* settlements are not allowed, allowing them brings two type of effects. First, it can increase the entry of generic firms due to higher rents. Specifically, in this model, these settlements induce G to develop a generic drug and file an ANDA to compete with drug A more often in the antecipation to receive the license fee from O. Second, generic entry may make the brand pharmaceutical shift preemptively the patent decision from developing the new drug B towards protecting the brand drug A. I restrict the attention to the case where, under *pay-for-delay* settlements, the only equilibrium dispute is settlement, by

introducing the assumption that follows.

Assumption 1: $F_S < F < F_S^{PD}$ and $K > C$

The purpose of the Assumption 1 is to narrow down the analysis and identify the role that *pay-for-delay* settlements are having in generic entry and on the innovation decision of brand pharmaceuticals. By comparing the thresholds F_S and F_S^{PD} one concludes that this region always exists. It first states that, under *regular* settlements, G does not enter market A if protection falls in any dispute region, but does so under *pay-for-delay* settlements. This way it simplifies the analysis and captures the entry effect only in the *pay-for-delay* setting. Secondly, Assumption 1 says that the settlement cost C must not be extremely high, meaning that under regular settlements O's reaction to entry can be any kind of dispute, settlement or trial. However, O never goes to trial if G enters under *pay for-delay* settlements (since $K > C$ implies that $\alpha_T^{PD} > 1$).

Under *regular* settlements, G only enters in regions $AcAc$, SAc and TAc. In the remaining regions G never enters, so protection of drug A, A_1, is valueless and O develops drug B. Likewise, in region $AcAc$ O chooses to develop drug B since a minor protection improvement of drug A is valueless. In regions SAc and TAc, the decision to invest in either strategy depends on the comparison of strategies' payoffs. O chooses to protect drug A if $\pi^M \geq \pi^D + R_A + \pi^B - R_B$, i.e., if the market for drug A is profitable enough and worth to protect, otherwise it always develops drug B.

Under *pay-for-delay* settlements, G files an ANDA for protection levels below θ_S^{PD} and O is willing to settle. In the case G enters in both cases and O settles, O's decision to develop the new drug B is based on the protection improvement of patent A_1.

If drug B is a very profitable drug or if improving protection involves high expenditures in R&D, then protection must be improved significantly in order to be preferred to the new drug. If protection levels remain similar enough, then drug B is more likely to be developed. When patent protection α_1 is high enough to deter G's entry, but α_0 is not and G enters if O develops drug B, then O chooses A_1 for a level of patent protection α_0 low enough, i.e., below the threshold level Θ. Also, in region SAc, G enters partially for protection level α_1 low enough and it that case, O chooses A_1 above a certain level of protection α_1, Ω. Otherwise, if α_1 is high enough and deters G from entering, O chooses to improve protection of drug A, such as in *regular* settlements.

So, on one hand, under *regular* settlements O improves protection of drug A in SAc and TAc regions if drug is initially weakly protected and is able to deter generic's entry. To do so, drug A must be profitable enough as well. Under pay-for-delay, O chooses both new drug B and protection of drug A (patent A_1) in region SAc under *pay-for-delay* settlements, deciding to protect when protection improvement is significant enough. So, the entry effect reduces the area where O improves protection of drug A.

On the other hand, under *regular* settlements O only chooses B in the dispute region where is able to deter entry (regions SS, TS, and TT), whereas under *pay-for delay* settlements this is not so, O decides between drug B and more protection of drug A. So, the entry effect increases the area where O improves protection of drug A exactly for the purpose of deterring this entry. The two effects go in opposite directions in making patent decisions. Also for some parameters, in region SS, firm G enters and O choose A_1 for an sufficiently attractive market A. I denote this region as $D = \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2$, where $\Upsilon = \theta_S^{PD} - \Psi$.

Therefore, i ask: when is the case that O invests more often in improving protection of drug A under *pay-for-delay* settlements with respect to invest in new drugs? Proposition 3..1 identifies when O invests in drug A more frequently under *pay-for-delay* settlements than under *regular* settlements. To simplify the notation, let's denote also $L \equiv \pi^M - \pi^D - R_A + R_B$, $N \equiv C + F - R_A + R_B$ and $S \equiv C + F + \pi^D$ and $M \equiv 2C + 2F + \pi^D$. I also abstract from the case where $\pi^B > L$, since in that case O always invests in drug B in both types of settlements and no conclusions are withdrawn from it.

Proposition 3..1 *Under Assumption 1, O invests more frequently in A under pay-for-delay settlements:*

i) If $\pi^M < S$: and $\pi^B < L - F - C$ and $\Omega < \theta_S^{PD} + \frac{\theta_S^{PD}(1-\theta_S^{PD}) - \alpha_S(1-\alpha_S) - D}{\alpha_S}$; and $L - F - C < \pi^B < N$, and $\theta_S^{PD} + \alpha_S - \frac{D}{\theta_S^{PD} - \alpha_S} < 1$.

ii) If $S < \pi^M < M$: and $\pi^B < L - F - C$ and $\Omega < \theta_S^{PD} + \frac{\theta_S^{PD}(1-\theta_S^{PD}) - \alpha_S(1-\alpha_S) - D}{\alpha_S}$; and $L - F - C < \pi^B < N$ and $\theta_S^{PD} + \alpha_S - \frac{D}{\theta_S^{PD} - \alpha_S} < 1$; and $N < \pi^B < L$ and $\Theta > \frac{\alpha_S(1-\alpha_S) - D}{(1-\theta_S^{PD})}$.

iii) If $\pi^M > M$: and $\pi^B < N$ and $\Omega < \theta_S^{PD} + \frac{\theta_S^{PD}(1-\theta_S^{PD}) - \alpha_S(1-\alpha_S) - D}{\alpha_S}$; and $N < \pi^B < L - C - F$ and $\Theta > \frac{\alpha_S(\Omega - \alpha_S) - D}{(1-\theta_S^{PD})} + \alpha_S$; and $L - C - F < \pi^B < L$ and $\Theta > \frac{\alpha_S(1-\alpha_S) - D}{(1-\theta_S^{PD})}$.

So indeed, even for attractive of new drugs, *pay-for-delay* settlements can alter the decision of brand pharmaceuticals and favour investment in protecting existing drugs. If market for drug A is sufficiently attractive, then this protection effect is even

stronger because of the entry effect that *pay-for-delay* settlements bring. This type of settlements is, then, introducing an incentive to undergo new drugs more frequently if the prospective market is more attractive than before, foregoing other drugs that, although less profitable, under *regular* settlements would possibly be invested in.

One empirically testable prediction is to expect that brand firms doing *pay-for-delay* settlements hold relatively smaller portfolios of new drugs (New Molecular Entities) than firms that do not perform these settlements. Also, one empirical prediction of this analysis is that, all else equal, it is more likely to observe *pay-for-delay* settlements in markets with lower entry costs, or higher profits.

Consumer welfare analysis I evaluate the consumer welfare implications of allowing *pay-for-delay* settlements. When Proposition 3..1 holds, consumer welfare under *pay-for-delay* settlements changes with respect to the consumer welfare under *regular* settlements. Since no entry exists in dispute regions, under *regular* settlements there is no change in consumer surplus of market A, which is the monopoly consumer surplus. Even with generic entry under *pay-for-delay*, O pays E in order not to compete. Hence, the consumer welfare analysis is focused on the consumer surplus created by market B.

As long as the protection effect of market A prevails under the conditions described in the proposition 3..1, drug B is relatively less atractive to firm O, which means that consumers are deprived of a new drug. The consumer welfare change the balance between the consumer surplus in market B, denoted as CS^B, that is gained when generics enter but protection improvement is not significant and O decides to develop a new drug B and the consumer surplus lost in market B due to a switch in O's strategy

to improve intermediate protection of drug A to prevent E from entering market A. For exposition purposes, I focus on one case of proposition 3..1. If $\pi^M > M$ and $N < \pi^B < L - C - F$, O invests more frequently in patent A_1 if $\Theta > \alpha_S + \frac{\alpha_S(\Omega-\alpha_S)-D}{(1-\theta_S^{PD})}$. In that case, Consumer Welfare Change (CWC):

$$CWC \ = CS^B[\alpha_S(\Omega - \alpha_S) - D - (\Theta - \alpha_S)(1 - \theta_S^{PD})]$$

The first term is the consumer access to market B and the second term is the loss in market B due to a protection effect that made O change its patenting strategy. So the welfare change is negative as long as the protection effect prevails, and positive otherwise, following the results of proposition 3..1. Consumers will only have the same access to a new drug under *pay-for-delay* settlements as under *regular* settlements when the profit of drug B is higher than before.

3.6 Example with Cournot competition

As an example I illustrate the results of Proposition 3..1 ii), when $S < \pi^M < M$, that is, the market A is sufficiently attractive. Consider inverse demand functions $p_A = 5 - q_A$ and $p_B = b - q_B$ of markets A and B, respectively. Assuming no production costs, profits are $\pi^M = \frac{25}{4}, \pi^D = \frac{25}{9}$ in market A. Profit in market B is $\pi^B = \frac{b^2}{4}$. Consider costs $R_A = 0.3$, $R_B = 0.8$, $F = 1.7$, $C = 1$ and $K = 1.2$. As B is assumed to be profitable, it must be that $b^2 > 3.2$. Also, in this case, $L = 3.97$, $L - F - C = 1.27$ and $N = 3.2$.

The thresholds for accommodation, settlement and trial are $\alpha_S = 0.35$ and $\alpha_T = 0.576$, when *pay-for-delay* settlements are not allowed. G does not enter in

dispute regions, since entry threshold $\theta_S = 0.34 < \alpha_S$. When *pay-for-delay* settlements are allowed, G enters for $\alpha < \theta_S^{PD} = 0.46$, $\alpha = \alpha_0, \alpha_1$ and O's reaction is to settle.

For different sizes of market B, it is possible to observe the trend of drug A protection. In particular, by increasing $B's$ market size (hence profitability), the protection of drug A remains more frequent under *pay-for-delay* settlements than under *regular* settlements.

When $b = 2$ (i.e., $\pi^B \leq L - F - C$) O invests more frequently in improving protection of drug A since $\Omega = 0.37$, while the relevant threshold is 0.46. When $b = 3$ (so that $L - F - C < \pi^B \leq N$), O still invests more frequently in A_1 as $\theta_S^{PD} + \alpha_S < 1$. When $b = 3.6$ (so that $N < \pi^B \leq L$), O invests more frequently in patent A_1 although $\Theta = 0.44$ is lower than the threshold 0.46. Finally, when $b = 3.7$ (under the same interval $N < \pi^B \leq L$), O invests less frequently in patent A_1 as Θ decreases 0.38 lower than the threshold 0.46.

The areas of improvement of drug A are calculated in Table 1, where RA_1 is the area under *regular* settlements and $PD_{A_1}$is the area under *pay-for-delay* settlements, for the different sizes of market for drug B.Indeed the regions have an area of 0.28 under pay-for-delay versus 0.23 under *regular* settlements. Fixing all other things constant, the A_1 region in the *pay-for delay* case is reduced as b increases. This means that O has the same relative preference in producing new drug B under *pay-for-delay* settlements and under *regular* settlements only when $b = 3.7$.

R_{A_1}	$PD_{A_1}(b=2)$	$PD_{A_1}(b=3.5)$	$PD_{A_1}(b=3.6)$	$PD_{A_1}(b=3.7)$
0.23	0.28	0.25	0.24	0.21

Table 1: Area where A_1 is chosen under the two regimes

Figures 3..3 and 3..4 depict the regions of dispute, G's entry decision, as well as O's patenting decision when $b = 2$, under both regimes. Comparing Figure 3..3, under *regular* settlements, with Figure 3..4, under *pay-for-delay* settlements, one can observe that *pay-for-delay* settlements induce more entry and a change in O's patenting strategy from developing new drugs to improve protection of the existing drug.

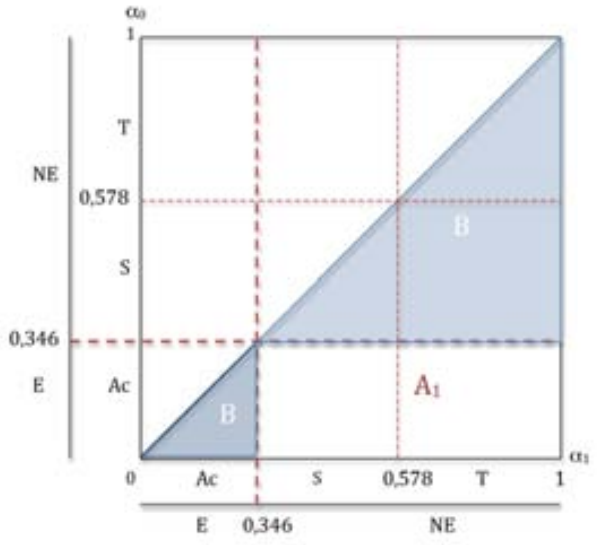

Figure 3..3: Regular Settlements

Additionally, in Table 2 I calculate the consumer welfare change, which is provided in Table 2 and given by:

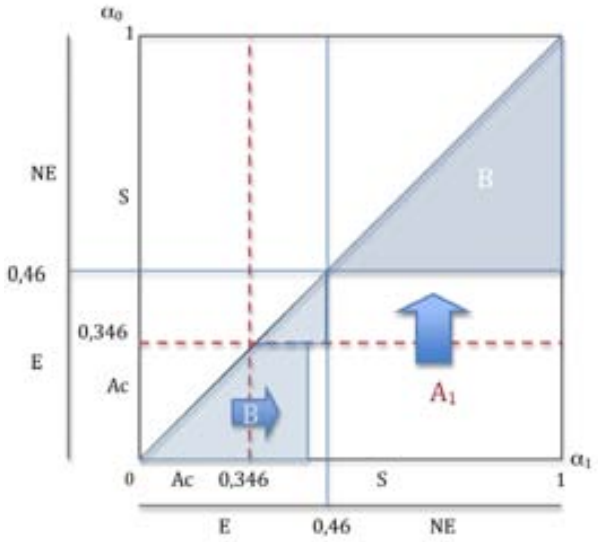

Figure 3..4: Pay-for-delay settlements

	$PD_{A_1}(b=2)$	$PD_{A_1}(b=3.5)$	$PD_{A_1}(b=3.6)$	$PD_{A_1}(b=3.7)$
CWC	-0.026	-0.032	-0.016	$0.03\,4$

Table 2: Consumer welfare change depending on size of market B

As long as the strategy of protection of drug A prevails, the consumer surplus falls, and this fall is greater for new drugs that create middle sized markets. This is so since O's patent decision remains unchanged but consumers are deprived of an increasingly important market B. Consumer welfare improves with respect to *regular* settlements again only when market B is attractive enough, i.e., for $b = 3,7$.

3.7 Discussion

In this section I briefly discuss the robustness of my results to some extension of the model.

Generic entry under *regular* settlements: The analysis has been focused on the clearest case of no entry under dispute in *regular* settlements in order to show the relevant effects that *pay-for-delay* settlements bring to the patent decision of brand pharmaceuticals. If it is the case that market A is sufficiently large (or entry cost is low enough) then it may attract generic entry under both settlement regimes. In this case, the entry effect that I identified becomes stronger. Generics continue to file ANDA applications more frequently in the perspective of receiving a generous license fee from brand pharmaceuticals in order not to compete.

In the model, taking into account the relevant thresholds upon which O decides, the qualitative results withdrawn in Proposition 3..1 remain the same, that is, there continues to exist cases where entry of generic G makes O prefer to protect drug A more often, in detriment of the new drug B. With respect to the consumer welfare change, the analysis must account for the loss in consumer surplus in market A as well, since if G enters, A_1 is chosen under *regular* settlements and both firms share the market in duopoly, a *pay-for-delay* settlement eliminates competition and deter consumers from having access to a cheaper drug A that otherwise regular settlements or litigation would allow.

Protection of drug A increases its profits: In the current framework, I have assumed that protection of drug A provides only judicial benefits. The analysis can consider that holding patent A_1 also provides higher profits in market A. There are cases where the profit increases through the launch of variations of the same drug with new dosages, or the introduction of a different type of tablet that allows a different body absorption rate, or an extension in the variety of products associated to the same

active ingredient (namely from pills to cream/pomade). There are two main effects that arise: O is willing to enter in dispute more often and G is willing to file an ANDA more often than if drug A did not face an increase in its profits.

The results attained in Proposition 3..1 continue to hold. O prefers to invest in A_1 more often under pay for delay settlements.The intuition is that, being O the incumbent of a higher profit market, it has the incentives to increase efforts in protecting the monopolist position and uphold its patent. In this case the dispute region is larger due to an decrease in the accommodation-settlement threshold, G has more incentives to enter due to a prospective higher profit, and O's settlement payoff under *pay-for-delay* settlement is higher than before in case it develops A_1.

When no kind of settlements are allowed: the benchmark case would only have trial as a possible reaction of O with respect to G's entry in market A. The analysis would obtain even stronger results since the generic firm G enters less often under a trial reaction than under *regular* settlements. So, the entry effect and protection effects increase even more when *pay-for-delay* settlements are allowed.

Portfolio Protection Effect from developing drug B: I considered that patenting drug B can produce a portfolio protection effect and help protect market A as well.[9] Although in the pharmaceutical industry this protection effect is not easy to identify, while in the electronics and semi-conductor sectors and where patents have been identified as bargaining chips in settlement negotiations (Hall and Ziedonis, 2001), I could consider a case where patent B also provides some protection to the portfolio. The qualitative results withdrawn in the analysis would not change however. The main

[9]For more on portfolio protection effect of patents, see Parchomovski and Wagner (2005).

effect is that there would be different thresholds for accommodation settlement and trial on behalf of firm O, as well as entry, but the protection effect of drug A would still prevails in many cases with *pay-for-delay* settlements and the analysis would retrieve the same result.

3.8 Conclusion

In this paper I introduce a framework where an brand drug firm that faces the potential entry of a generic drug firm before patent expiration. Entry can occur under the Paragraph IV certification of the Hatch-Waxman Act. I use patent strength to measure the incumbent's probability of having its patent upheld and found valid in court. The incumbent, can either sue or accommodate the entrant firm, where after suing the parties either go to trial or settle.

In deciding on what innovations to follow, I model the choice of a brand pharmaceutical that decides to allocate resources in either improving protection of a drug for which it is the patentholder and incumbent, or to develop a new patent that allows to be the monopolist in the market of a new drug B, taking into account the future possibility of entry of a generic drug firm in the market of the already existing drug A.

I identify that, in some cases, *pay-for-delay* settlements to resolve patent disputes induce generic firms to enter more often in brand drug markets and to shift brand firms' patent decisions into more protection of existing drug, delaying the development or launch of new drugs or prefering to only develop drugs that present substantial profitability. So, *pay-for-delay* settlements may be already inducing more investment on behalf of generics to enter in existing markets but, at the same time, brand pharma-

ceuticals are investing less often in new drugs due to this entry threat.

This paper offers a new explanation on the role that *pay-for-delay* settlements are playing in the pharmaceutical sector and in the decreasing trend of new drugs being lauched to the market. This is one first approach to the problem. I have only considered the patent portfolio is just composed of two patents, and this analysis could be extended to a portfolio with more patents. Also, a new approach to the problem can be to consider licensing contracts not only involving a fixed compensation, but also in the form of royalties.

3.9 Appendix

Proof of Lemma 3..1 Since payoffs only change due to different protection levels, I solve for any α, $\alpha \in \{\alpha_0, \alpha_1\}$. I first analyze whether O wants to settle or not. When $Ac_O \geq T_O \Leftrightarrow \pi^D \geq \alpha\pi^M + (1-\alpha)\pi^D - K \Leftrightarrow K \geq \alpha(\pi^M - \pi^D)$, the reservation value for any settlement negotiation is the duopoly profit, since if negotiations were to fail, O chooses to go to trial and accommodates. As it is a negotiation with no risk of trial, there is no bargaining surplus (zero-sum game). The settlement process is not mutually profitable compared to accommodating, so O always accommodates.

If $K < \alpha(\pi^M - \pi^D)$, then trial is a "back-up solution" for O. In case of settlement, the optimal license fee is the solution to:

$$\max_{L_\alpha} [S_O - T_O]^{0.5} [S_G - T_G]^{0.5}$$

where 0.5 is the bargaining power of O. The optimal license is:

$$L_{\alpha}^{*} = 0.5\alpha\pi^{M}$$

So, if O improves protection of drug A its payoff is $S_O = \pi^D + L_\alpha - R_A$ and if develops new drug B is $S_O = \pi^D + L_{\alpha_0} + \pi^B - R_B$. G's payoff under settlement is $S_G = \pi^D - L$. Since, in the case O develops drug B, $\pi^B - R_B$ exists independently if O goes to trial or settles, it cancels out in the comparison of payoffs. Hence, the general condition under which O prefers to settle than to go to trial is:

$S_O \geq T_O \Leftrightarrow \pi^D + 0.5\alpha\pi^M - R_A - C \geq \alpha\pi^M + (1-\alpha)\pi^D - R_A - K \Leftrightarrow C \leq K - 0.5\alpha(\pi^M - 2\pi^D)$

In terms of patent protection, O accommodates for $\alpha \leq \frac{K}{\pi^M - \pi^D} \equiv \alpha_S$,settles for $\alpha \in (\alpha_S, \alpha_T]$, where $\alpha_T \equiv \frac{2(K-C)}{\pi^M - 2\pi^D}$ and goes to trial for higher protection levels. The thresholds are the same independently of the patenting decision. In order for the 3 regions to exist, the following conditions must hold:

$$K - \frac{\pi^M - 2\pi^D}{2} \leq C < 0.5K\frac{\pi^M}{\pi^M - \pi^D}$$

Proof of Lemma 3..2 As the thresholds for accomodation, settlement and trial are the same independently of what O patents, I solve for any α, $\alpha \in \{\alpha_0, \alpha_1\}$. If patent strength $\alpha \in (0, \alpha_S]$, then O accommodates and G always enters since $\pi^D - F \geq 0$. However, if $\alpha \in (\alpha_S, 1]$, then O enters in dispute and this influences G's entry decision.

1. If strength $\alpha \in (\alpha_S, \alpha_T]$, O settles in case G enters. G enters if profits are positive:

$S_G \geq 0 \Leftrightarrow \pi^D - 0.5\alpha\pi^M - F \geq 0 \Leftrightarrow \alpha \leq \frac{2(\pi^D - F)}{\pi^M} = \theta_S$

This entry region exists only if: $\theta_S > \alpha_S \Leftrightarrow F < \pi^D - 0.5K\frac{\pi^M}{\pi^M - \pi^D} = F_S$

Otherwise, G never enters in the settlement region.

2. If strength $\alpha \in (\alpha_T, 1]$, O goes to trial in case G enters. G enters if profits are positive:

$T_G \geq 0 \Leftrightarrow (1-\alpha)\pi^D - K - F \geq 0 \Leftrightarrow F \leq (1-\alpha)\pi^D - K \Leftrightarrow \alpha \leq 1 - \frac{K+F}{\pi^D} = \theta_T$

This entry region exists only if $\theta_T > \alpha_T \Leftrightarrow F < \pi^D + C\frac{2\pi^D}{\pi^M - 2\pi^D} - K\frac{\pi^M}{\pi^M - 2\pi^D} = F_T$

Otherwise, G never enters in the trial region.

Proof of Lemma 3..3 If O is the only one in the market independently of what might patent, then B is chosen if: $B \geq A_1 \Leftrightarrow \pi^M + \pi^B - R_B \geq \pi^M - R_A \Leftrightarrow \pi^B - R_B \geq -R_A$, which always holds.So, I do not refer this case in any region.

AcAc *When* $\alpha \in [0, \alpha_S]$, $\alpha = \alpha_0, \alpha_1$, O develops B.

SS *When* $\alpha \in (\alpha_S, \theta_S]$, $\alpha = \alpha_0, \alpha_1$, B is chosen if:

$\pi^D + \pi^B - R_B + L^*_{\alpha_0} - C \geq \pi^D - R_A + L^*_{\alpha_1} - C$

$\Leftrightarrow \alpha_0 \geq \alpha_1 - \frac{2(\pi^B - R_B + R_A)}{\pi^M} \Leftrightarrow \alpha_1 - \alpha_0 \leq \Psi$. Otherwise chooses A_1.

When $\alpha_0 \in (\alpha_S, \theta_S]$ and $\alpha_1 \in (\theta_S, 1]$**,** B is chosen if:

$\pi^D + \pi^B + L^*_0 - R_B - C \geq \pi^M - R_A$

$\Leftrightarrow \alpha_0 \geq \frac{2(\pi^M - \pi^D - R_A - \pi^B + R_B + C)}{\pi^M} \Leftrightarrow \alpha_0 \geq \psi$. Otherwise A_1.

TT *When* $\alpha \in (\alpha_T, \theta_T]$ B is chosen if:

$\alpha_0\pi^M + (1-\alpha_0)\pi^D + \pi^B - R_B - K \geq \alpha_1\pi^M + (1-\alpha_1)\pi^D - R_A - K$

$\Leftrightarrow \alpha_0 \geq \alpha_1 - \frac{\pi^B - R_B + R_A}{\pi^M - \pi^D} \Leftrightarrow \alpha_1 - \alpha_0 \leq \varpi$. Otherwise, chooses A_1.

When $\alpha_0 \in (\alpha_T, \theta_T]$ and $\alpha_1 \in (\theta_T, 1]$**,** B is chosen if:

$\Leftrightarrow \alpha_0(\pi^M - \pi^D) \geq \pi^M - \pi^D - R_A - \pi^B + R_B + K$

$\Leftrightarrow \alpha_0 \geq 1 - \frac{\pi^B - R_B + R_A - K}{\pi^M - \pi^D} \Leftrightarrow \alpha_0 \geq 1 + \alpha_S - \varpi$

SAc *When* $\alpha_0 \in [0, \alpha_S]$ and $\alpha_1 \in (\alpha_S, \theta_S]$**,** B is chosen if:

$\pi^D + \pi^B - R_B \geq \pi^D - R_A + L^*_{\alpha_1} - C \Leftrightarrow \alpha_1 \leq \frac{2(\pi^B - R_B + R_A + C)}{\pi^M} \Leftrightarrow \alpha_1 \leq \gamma$

When $\alpha_0 \in [0, \alpha_S]$ and $\alpha_1 \in (\theta_S, 1]$, B is chosen if:

$\pi^D + \pi^B - R_B \geq \pi^M - R_A \Leftrightarrow \pi^B - R_B \geq \pi^M - \pi^D - R_A$. Otherwise A_1.

TAc *When* $\alpha_0 \in [0, \alpha_S]$ and $\alpha_1 \in (\alpha_T, \theta_T]$**,** B is chosen if:

$\Leftrightarrow \pi^D + \pi^B - R_B \geq \alpha_1 \pi^M + (1 - \alpha_1)\pi^D - R_A - K \Leftrightarrow \alpha_1 \leq \frac{R_A + K + \pi^B - R_B}{\pi^M - \pi^D} \Leftrightarrow \alpha_1 \leq \alpha_S + \varpi$

When $\alpha_0 \in [0, \alpha_S]$ and $\alpha_1 \in (\theta_T, 1]$**,** B is chosen if:

$\pi^B - R_B \geq \pi^M - \pi^D - R_A$. Otherwise chooses A_1.

TS *When* $\alpha_0 \in (\alpha_S, \theta_S]$ and $\alpha_1 \in (\alpha_T, \theta_T]$**,** B is chosen if:

$\pi^D + \pi^B - R_B + L^*_{\alpha_0} - C \geq \alpha_1 \pi^M + (1 - \alpha_1)\pi^D - R_A - K$

$\Leftrightarrow \alpha_0 \geq \alpha_1 2(1 - \frac{\pi^D}{\pi^M}) - \frac{2(\pi^B - R_B + R_A + K - C)}{\pi^M} \Leftrightarrow \alpha_1 \lambda - \alpha_0 \leq \Gamma$. Otherwise chooses A_1.

When $\alpha_0 \in (\alpha_S, \theta_S]$ and $\alpha_1 \in (\theta_T, 1]$**,** B is chosen if:

$\pi^D + \pi^B - R_B + L^*_{\alpha_0} - C \geq \pi^M - R_A$

$\Leftrightarrow \alpha_0 \geq 2 - \frac{2(\pi^D + R_A - C + \pi^B - R_B)}{\pi^M} \Leftrightarrow \alpha_0 \geq \psi$. Otherwise A_1.

Proof of Lemma 3..4 Once more, I solve for any α, $\alpha \in \{\alpha_0, \alpha_1\}$. I follow the same procedure as in the proof of Lemma 3..1. Let's see if O wants to settle or not. If $K \geq \alpha(\pi^M - \pi^D)$, the negotiation has no risk of trial, since no bargaining surplus exists as it is a zero-sum game. Settlement is not a mutually profitable compared to accommodating, so O always accommodates. If, $K < \alpha(\pi^M - \pi^D)$, trial is a "back-up solution" for O. Under settlement, the optimal license fee is the solution to:

$$\max_L [S_O^{PD} - T_O]^{0.5}[L - T_G]^{0.5}$$

The optimal license fee is:

$$L_\alpha^{PD*} = 0.5(1-\alpha)\pi^M$$

O's payoff is $S_O = \pi^M - L_{\alpha_1}^{PD*}$ if improves protection of drug A and $S_O = \pi^M - L_{\alpha_0}^{PD*} + \pi^B - R_B$ if develops patent B. O prefers to settle over going to trial if $S_O \geq T_O \Leftrightarrow C \leq 0.5(1-\alpha)(\pi^M - 2\pi^D) + K$. In terms of patent protection, O accommodates for $\alpha \leq \frac{K}{\pi^M - \pi^D} = \alpha_S$,settles for $\alpha \in (\alpha_S, \alpha_T^{PD}]$, where $\alpha_T^{PD} = 1 + \frac{2(K-C)}{\pi^M - 2\pi^D}$ and goes to trial for higher protection levels. The thresholds are the same independently of the patenting decision. In order for the 3 regions to exist, the following conditions must hold:

$$K < C < 0.5(\pi^M - 2\pi^D) + 0.5K\frac{\pi^M}{\pi^M - \pi^D}$$

Proof of Lemma 3..5: I again solve for a general protection level α, $\alpha = \alpha_0, \alpha_1$. If $\alpha \in (0, \alpha_S]$, O accommodates and G always enters since $\pi^D - F \geq 0$. If $\alpha \in (\alpha_S, 1]$ O

settles and this influences G's entry decision. If strength $\alpha \in (\alpha_S, \alpha_T^{PD}]$, O settles and G enters if profits are positive:

$$S_G \geq 0 \Leftrightarrow L^{PD*} - F \geq 0 \Leftrightarrow 0.5(1-\alpha)\pi^M - F \geq 0$$

$$\alpha \leq 1 - \frac{2F}{\pi^M} = \theta_S^{PD}$$

This entry region exists only if $\theta_S > \alpha_S \Leftrightarrow F < 0.5\pi^M - 0.5K\frac{\pi^M}{\pi^M - \pi^D} = F_S^{PD}$. Otherwise, G never enters in the settlement region.

Proof of Lemma 3..6: As in lemma 3, when G does not enter independently of O's patent strategy, it always chooses patent B. Hence, I will not refer to this case in any of the regions.

SS *When* $\alpha \in (\alpha_S, \theta_S^{PD}]$, $\alpha = \alpha_0, \alpha_1$, B is chosen if:

$\Leftrightarrow \pi^M + \pi^B - L_{\alpha_0}^{PD*} - R_B - C \geq \pi^M - L_{\alpha_1}^{PD*} - R_A - C$

$\Leftrightarrow \alpha_0 \geq \alpha_1 - \frac{2(\pi^B - R_B + R_A)}{\pi^M} \Leftrightarrow \alpha_1 - \alpha_0 \leq \Psi$. Otherwise, chooses A_1.

When $\alpha_0 \in (\alpha_S, \theta_S^{PD}]$ and $\alpha_1 \in (\theta_S^{PD}, 1]$**,** B is chosen if:

$\pi^M + \pi^B - L_{\alpha_0}^{PD*} - R_B - C \geq \pi^M - R_A$

$\Leftrightarrow \alpha_0 \geq 1 - \frac{2(\pi^B - R_B + R_A - C)}{\pi^M} \Leftrightarrow \alpha_0 \geq \Phi$. Otherwise chooses A_1.

SAc *When* $\alpha_0 \in [0, \alpha_S]$ and $\alpha_1 \in (\alpha_S, \theta_S^{PD}]$**,** B is chosen if:

$\pi^D + \pi^B - R_B \geq \pi^M - L_{\alpha_1}^{PD*} - R_A - C$

$\Leftrightarrow \alpha_1 \geq 1 - \frac{2(\pi^B - R_B + R_A + C + \pi^D)}{\pi^M} \Leftrightarrow \alpha_1 \geq \Omega$. Otherwise chooses A_1.

When $\alpha_0 \in [0, \alpha_S]$ and $\alpha_1 \in (\theta_S^{PD}, 1]$, B is chosen if:

$\pi^D + \pi^B - R_B \geq \pi^M - R_A \Leftrightarrow \pi^B - R_B \geq \pi^M - \pi^D - R_A$. Otherwise, A_1.

TS *When* $\alpha_0 \in (\alpha_S, \alpha_T]$ and $\alpha_1 \in (\alpha_T, \theta_T^{PD}]$**,** B is chosen if:

$\Leftrightarrow \pi^M + \pi^B - L^{PD*} - R_B - C \geq \alpha_1 \pi^M + (1-\alpha_1)\pi_D^D - R_A - K$

$\Leftrightarrow \alpha_0 \geq \alpha_1 \frac{2(\pi^M - \pi^D)}{\pi^M} - 1 - \frac{2(\pi_M^B - R_B + R_A + K - C - \pi^D)}{\pi^M} \Leftrightarrow \alpha_1 \lambda - \alpha_0 \leq \delta$. Otherwise, A_1.

When $\alpha_0 \in (\alpha_S, \alpha_T]$ and $\alpha_1 \in (\theta_T^{PD}, 1]$**,** B is chosen if:

$\Leftrightarrow \pi^M + \pi^B - L^{PD*} - R_B - F_S \geq \pi^M - R_A$

$\Leftrightarrow \alpha_0 \geq 1 - \frac{2(\pi^B - R_B + R_A - C)}{\pi^M} \Leftrightarrow \alpha_0 \geq \Phi$

Proof of Proposition 3..1 G does not enter in any dispute region under *regular* settlements and enters for $\alpha \leq \theta_S^{PD}$, $\alpha = \alpha_0, \alpha_1$ under *pay-for-delay* settlements. Under *regular* settlements, O chooses B in $AcAc$, SS, TS and TT regions, and A_1 in SAc and TAc regions. So, under regular settlements, the area where A_1 is chosen is given by:

$$R_{A_1} = \alpha_S(1 - \alpha_S)$$

Under *pay-for-delay* settlements, O chooses B in $AcAc$. In the SS and SAc regions, O may choose differently according to the parameters.

In region SS if G enters then O's choice between B and A_1 depends on how profitable market A is. $\Upsilon = \theta_S^{PD} - \Psi = 1 - \frac{2(F + \pi^M - R_B + R_A)}{\pi^M}$ is the level of protection α_0 when $\alpha_1 = \theta_S^{PD}$. Υ is only greater than α_S when $0.5(\pi^M - 2\pi^D) - R_A > \pi^B - R_B$. Overall, there are four possible cases:

1) $\alpha_S < \Omega < \Theta < \theta_S^{PD} \Leftrightarrow \pi^M > \pi^D + F + C + R_A + \pi^B - R_B$ and $\pi^B > F + C - R_A + R_B$

In Pay-for-delay settlements, the area A_1 is:

$PD_{A_1} = \Theta(1-\theta_S^{PD}) + \alpha_S(\theta_S^{PD} - \Omega) + \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2$

So, O chooses A_1 more times under *pay-for-delay* settlements if:

$PD_{A_1} > R_{A_1} \Leftrightarrow (\Theta - \alpha_S)(1-\theta_S^{PD}) + \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2 > \alpha_S(\Omega - \alpha_S) \Leftrightarrow$

$\Theta > \alpha_S + \frac{\alpha_S(\Omega-\alpha_S) - \frac{1}{2}(Max\{\Upsilon,\alpha_S\}-\alpha_S)^2}{(1-\theta_S^{PD})}$

2) $\alpha_S < \Omega < \theta_S^{PD} < \Theta \Leftrightarrow \pi^M > \pi^D + F + C + R_A + \pi^B - R_B$ and $\pi^B < F + C - R_A + R_B$

In Pay-for-delay settlements, the area A_1 is: $PD_{A_1} = \theta_S^{PD}(1-\theta_S^{PD}) + \alpha_S(\theta_S^{PD} - \Omega)$

So, O chooses A_1 more times under *pay-for-delay* settlements if:

$PD_{A_1} > R_{A_1} \Leftrightarrow (\theta_S^{PD} - \alpha_S)(1-\theta_S^{PD}) + \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2 > \alpha_S(\Omega - \alpha_S) \Leftrightarrow$

$\frac{(\theta_S^{PD}-\alpha_S)(1-\theta_S^{PD}) - \frac{1}{2}(Max\{\Upsilon,\alpha_S\}-\alpha_S)^2}{\alpha_S} + \alpha_S > \Omega.$

3) $\Theta < \theta_S^{PD} < \Omega \Leftrightarrow \pi^M < \pi^D + F + C + R_A + \pi^B - R_B$ and $\pi^B > F + C - R_A + R_B$ In Pay-for-delay settlements, the area A_1 is: $PD_{A_1} = \Theta(1-\theta_S^{PD}) + \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2$

So, O chooses A_1 more times under *pay-for-delay* settlements if:

$PD_{A_1} > R_{A_1} \Leftrightarrow \Theta > \alpha_S \frac{(1-\alpha_S) - \frac{1}{2}(Max\{\Upsilon,\alpha_S\}-\alpha_S)^2}{(1-\theta_S^{PD})}$

4) $\theta_S^{PD} < \Omega < \Theta \Leftrightarrow \pi^M < \pi^D + F + C + R_A + \pi^B - R_B$ and $\pi^B < F + C - R_A + R_B$ In Pay-for-delay settlements, the area A_1 is: $PD_{A_1} = \theta_S^{PD}(1-\theta_S^{PD}) + \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2$

So, O chooses A_1 more times under *pay-for-delay* settlements if:

$PD_{A_1} > R_{A_1} \Leftrightarrow \theta_S^{PD}(1-\theta_S^{PD}) + \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2 > \alpha_S(1-\alpha_S) \Leftrightarrow \theta_S^{PD} -$

$\alpha_S + \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2 > \theta_S^{PD^2} - \alpha_S^2$

$\Leftrightarrow \theta_S^{PD} - \alpha_S > (\theta_S^{PD} + \alpha_S)(\theta_S^{PD} - \alpha_S) - \frac{1}{2}(Max\{\Upsilon, \alpha_S\} - \alpha_S)^2 \Leftrightarrow \quad \theta_S^{PD} + \alpha_S -$

$\frac{\frac{1}{2}(Max\{\Upsilon,\alpha_S\}-\alpha_S)^2}{\theta_S^{PD}-\alpha_S} < 1.$

REFERENCES CITED

1. Aoki, R. and J.L. Hu, 1999, "Licensing vs. Litigation: The Effect of the Legal System on Incentives to Innovate," Journal of Economics and Management Strategy 8(1): 133–160.

2. Bessen, J. E. and M.J. Meurer, 2006, "Patent Litigation with Endogenous Disputes", American Economic Review 96(2): 77-81

3. Blind, K. and F. Köhler, 2010, "Claim Amendments as a Result of Strategic Patenting and as a Driver for Patent Value" Working Paper, Berlin University of Technology.

4. Cohen, W., A. Goto, A. Nagata, R. Nelson and J. Walsh, 2002, "R&D spillovers, patents and the incentives to innovate in Japan and the United States", Research Policy 31: 1349-1367

5. Crampes, C. and C. Langinier, 2002, "Litigation and Settlement in Patent Infringement Case", RAND Journal of Economics 33(2): 258-274

6. Federal Trade Commission Report, 2011, "Agreements Filed with the FTC under the Medicare Prescription Drug, Improvement and Modernization Act of 2003"

7. Federal Trade Commission Statement, 2009, "How Pay-for-delay Settlements Make Consumers and the Federal Government Pay More for Much Needed Drugs"

8. Hall, B. and R. Ziedonis, 2001, "The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979-1995", Rand Journal of Economics 32(1): 101-128.

9. Hemphill, C. , 2006, "Paying for Delay: Pharmaceutical Patent Settlement as a Regulatory Design Problem", New York University Law Review 81: 1553-1574

10. Kash, D. and W. Kingston, 2001 "Patents in a world of complex technologies", Science and Public Policy 28: 11-22

11. Lanjouw, J. O. and M. Schankerman, 2001, "Characteristics of Patent Litigation: A Window on competition", RAND Journal of Economics 32(1): 129-151

12. Lanjouw, J. O. and M. Schankerman, 2004, "Protecting Intellectual Property Rights: Are Small Firms Hanicapped?", The Journal of Law and Economics 47 (1): 45-74

13. Meurer, M. J., 1989, "The Settlement of Patent Litigation", RAND Journal of Economics 20(1): 77-91

14. Parchomovsky, G. and R. P. Wagner 2005 "Patent Portfolios" 154 University of Pennsylvania Law Review, 12-14

15. PhRMA statement on patent settlements, 2010, http://www.phrma.org/media/releases/phrma-statement-patent-settlements-0

16. Schrag, J., 2007, "Economics at the FTC: Pharmaceutical Patent Dispute Settlements and Behavioral Economics", Review of Industrial Organization 31(2): 85-105

17. Schweitzer, S., 2007, "Pharmaceutical Economics and Policy", Oxford University Press.

18. Shapiro, C., 2003, "Antitrust Limits to Patent Settlements", RAND Journal of Economics 34(2): 391-411

19. Willig, R. and J. Bigelow, 2004, "Antitrust Policy toward Agreements That Settle Patent Litigation" 49: Antitrust Bullettin 655

CHAPTER 4.

THE NEUTRALITY DEBATE UNDER COMPETITION BETWEEN INTERNET SERVICE PROVIDERS

4.1 Introduction

The access to internet is crucial to guarantee the access to one of the most extensive ways of global and local communication nowadays. To be online, a consumer must purchase an online connection service to a firm that provides the network to navigate and consult contents in the internet. Such firms are called Internet Service Providers (ISPs). The online market is composed of three types of players: Internet Service Providers, networks that work as platforms to connect users that wish to access contents online; content providers which are producers of contents displayed online; and users, who browse the contents.

This paper evaluates the long run investment incentives of ISPs, that is, how much they invest in network capacity, under two different network regimes: network neutrality and network discrimination. This analysis considers that ISPs compete for user subscription and that there is network congestion, i.e. the time users have to wait to access a content depends on the number of users online.

Although there are several interpretations on net neutrality, they all coincide on the concept that all internet traffic should be treated in an identical way. The interpretation that we use of network neutrality is that ISPs cannot impose restrictions on users' acess to contents by prioritizing traffic and favoring certain data packets over others (Schuett, 2010). This means that ISPs cannot offer a tiered service, a priority

lane, to content providers and all users must have access to contents in an equal way independently of the content they browse.

Higher network capacity is known to allow users to access internet in an easier and faster way. The importance of the incentives to innovate in network capacity on behalf of ISPs, that is, to upgrade the technology of their networks and improve the connectivity between users and contents, is one of the main points of the debate on network neutrality.

The net neutrality debate which has been for some time in the United States and only recently entered the agenda in Europe. The main players concern in the debate is, besides the equal access to internet on behalf of users, the incentives of firms to invest in innovation. While in the U.S. the debate concerns whether or not a resolution should impose neutrality in online service, in Europe the position of the European Comission has been not to prevent discrimination of contents, but rather to promote competition between ISPs in the European market to minimize the potential problems arising from discrimination (abbusive power in network management) and, at the same time, to be a driver of investment in network capacity. To the European regulators, providing users the freedom to choose ISP and to switch if they are not satisfied should limit unreasonable traffic management of ISPs, intervening only under situations of unacceptable degradation of services (Sluijs, 2010).

In the recent literature this topic has already been studied under different approaches. Hermalin and Katz (2007) follow a contractual approach by considering a monopolist ISP that charges different fees to CPs for different qualities of online connection, and CPs differ in the attractiveness of their content. The ISP has no information

on the type of CPs, so it offers a contract menu to screen CPs. Choi and Kim (2010), Cheng (2011), and Kramer and Wiewiorra (2009) are the first to incorporate network congestion following queuing theory in the net neutrality debate literature. However, they have only considered the possibility of monopoly setting.[1] The main contribution of our paper is to introduce ISP competition with internet congestion. This way we extend the approach of Cheng and of Choi and Kim. We adopt a simple congestion model that is able to deliver the same qualitative results as Choi and Kim's model, and apply it to a competition setting.

The model of internet and content service developed in our paper is a two-sided market where consumers have heterogeneous preferences with respect to asymmetric ISPs and CPs. ISPs and CPs compete for end-users in a duopoly. We assume multi-homing of CPs, that is, content providers can provide content to users through more than one ISP. There are other network structures over which contents are delivered to consumers, however we chose multi-homing since it is possible to capture the direct interaction between each content provider and each internet service provider.

Internet congestion is a crucial element to this analysis since otherwise the effects of a discriminatory regime on the shift of users between contents could not correctly be accounted for. Also, introducing competition among ISPs is an important contribution to the literature for the two following reasons. Firstly, although an ISP monopoly may be realistic in some geographic areas of lower population density and where the costs to

[1] Other authors consider ISPs competition but in different frameworks. For instance, Baake and Mitusch (2007) study ISPs Cournot and Bertrand competition when consumers face congestion externalities. However, in their paper, the authors do not consider the network discrimination problem: ISPs are not allowed to prioritize one of the contents. In another paper, Musacchio (2009) model competition between several ISPs. In this case, the focus in not on consumers externality due to congestion but on the value consumers attribute to CPs investment in content quality.

penetrate are higher, in urban areas the choice of ISP connection is a reality and ISPs compete to obtain user subscription. Secondly, the European legal framework is set such that the net neutrality debate is not regarded as being a problem, so introducing competition enables to have a grasp of the effects that can be occurring in the european market and possibly provide a policy instrument.

Our main result is that competition between networks provides lower investments in capacity when network discrimination is allowed. The result holds if ISPs can charge a high fee to CPs. Contrary to what happens in a monopolistic network, ISPs market size is not fixed but can be increased to the detriment of the competitor. Network discrimination harms part of the consumers, hence end-users migrate to the network who penalize them less. If networks have asymmetric capacities, this translates into a transfer of consumer from the larger to the smaller network. As a result, the smaller network has lower incentives to invest because network discrimination partially reduces the gap between the two networks without requiring to increase capacity. If the larger network invests in capacity, it can mitigate users outflow but this would reduce the revenue from the priority fee. However, the loss of consumer due to network discrimination can be compensated if the fee charged to CPs is high enough. The overall effect is that ISPs prefer to discriminate between content and have lower incentive to expand the capacity of their networks.

The rest of the paper is organized as follows. The model is described in Section 2 and a brief illustration of how network congestion works is provided. Section 3 shows how this work is related to the seminal paper of Choi and Kim (2010). In Section 4, we present the equilibrium , when there is duopoly competition among ISPs under a

neutral regime. This constitutes a benchmark for future comparisons. In section 5, we determine the equilibrium outcomes where network discrimination is allowed. In Section 6, we compare investments incentives between the two regimes described in the previous sections. Section 7 provides conclusions and some policy indications. All proofs are remitted to the Appendix.

4.2 The Model

Basic Model An online network provides internet users the access access to online contents. This internet acess is sold by ISPs. CPs are the producers and deliverers of online content to end-users. To deliver the content, CPs must use the networks provided by ISPs. The network structure that we assume throughout the paper is that each CP is directly associated to more than one ISP, called multi-homing, and that all consumers are single-homing, i.e., they choose only one ISP with which to connect to.[2] The market for internet access, as well as the market for contents, are duopolies. ISPs are denoted as ISP_A and ISP_B, and the CPs as CP_1 and CP_2.

Net Neutrality. ISPs connect users to CPs and may be allowed to control how content is delivered to internet users. Under net neutrality, ISPs are not allowed to treat content providers differently, i.e., the time users wait to access content is regardless of the content they browse, nor to charge any price to CPs for the service. Under network discrimination, each ISP is allowed to sell a priority service to one of the CPs and charge a fee f (two-tiered pricing). We assume this fee is attained by Nash bargaining, where

[2] Although in practice the possibility of multi-homing exists, it is not common in user behavior. The great majority of users purchases internet services from one provider only.

θ is the bargaining power of the ISP selling the priority service. The fee is an amount between the maximum willingness to pay of the more efficient CP and the maximum willingness to pay of the less efficient CP. A user that browses a priority content receives it ahead of any other user requesting a non-priority content. Therefore, the time users wait for the requested content varies and this affects their utility.

Users. There is a mass (1×1) of users with heterogeneous preferences regarding: (i) the two contents; and (ii) the two internet online services. Each user is described by (x, y), where $x \in [0, 1]$ denotes the preferred type of content and $y \in [0, 1]$ the preferred type of online service. A user pays a price p to the ISP. Each user values the content and the online service at $v > 0$, where v is always high enough to guarantee full market coverage. If a user is not able to access her preferred content, she faces a utility loss in the form of a transport cost $t > 0$ times the distance between the preferred and the browsed content. Similarly, if a user cannot connect to her preferred online service, the utility loss is given by a transport cost $s > 0$ times the distance between the preferred and the subscribed service.

Each user demands only one of the two CPs and one of the two ISPs, also known as single-homing. When deciding which content to browse, her utility is affected by the time she waits to receive the content. Such waiting disutility, w, is assumed to be independent on the content type and is determined by the degree of congestion in the network, that is, the number of users online. However, when each user decides which ISP to access to, she is cannot determine the exact network congestion since it is not possible to have that information before subscribing an online service and browsing a

content. So, to estimate her waiting disutility, she uses the network capacity of the ISP, the perception of the number of connected users, and the possible prioritization of the content as congestion proxies.

In practice, a user can observe the network capacity of each ISP and which content is prioritized, but she does not know how many users in her building, or neighborhood, are also browsing the same content through the same ISP. Since the total number of users determines the congestion in a network, she can only observe the exact disutility of browsing a content once subscribed to an ISP. Then, she may stick to her dicision to browse the content she prefers most or switch to the other content. However, once subscribed to an ISP she cannot switch to the other ISP due to high switching costs (e.g. a permanence obligation).

Users' demand for contents is captured by demand intensity parameter λ, which is the same to all the users. It is a measure of the time spent browsing a content or the number of clicks in the content's page. Hence, the utility of a user (x, y) browsing content type $\bar{x}$ with an online service type $\bar{y}$ is

$$u_{x,y} \equiv v - t|x - \bar{x}| - s|y - \bar{y}| - w - p.$$

ISPs. The competition setup between ISP_A and ISP_B is the Hotelling model. They offer two different types of services and set prices p_j, $j = A, B$. For example, one ISP can provide an internet service including free movies and the other ISP including free sport programs. Ultimately one of them could provide Internet and television access, and the other Internet and phone calls access.

We assume the ISPs are located at the extreme points of the end-users' preference

line. ISP_A offers a service type $y = 0$ while ISP_B type $y = 1$. Hence, user y faces a transport cost sy when subscribing to ISP_A and $s(1-y)$ when subscribing to ISP_B.

Each ISP is endowed with a network capacity of μ_j, $j = A, B$. In the short run, μ_j is fixed. In the long run, it is endogenous to the model. A higher network capacity is associated with shorter waiting time to receive a content. In order to reduce the number of cases to analyze, we assume that ISP_A is the large network: $\mu_A > \mu_B$.

If an ISP is allowed to discriminate, each one decides which content to prioritize and charges a fee f_j. Managing data traffic does not imply any cost. We assume the costs to provide internet services are sunk, so ISP_j profits are given by

$$\pi_j = \sigma_j p_j + f_j$$

where σ_j denotes the market share of ISP_j.

CPs. The competition setup between CP_1 and CP_2 is also the Hotelling model. compete to deliver contents to end-users. Users only browse one of the two contents.[3] CP_1 is located at point 0 and CP_2 at point 1 of users' content preference line. A user located at x incurs a transport cost tx by browsing CP_1, while $t(1-x)$ by browsing CP_2.

Each CP_i, $i = 1, 2$, adopts a business model that provides contents without receiving payment from users. The revenues are exclusively from advertising, obtaining r_i from advertisers for each time a content is browsed.[4] The cost of providing a content is constant, c_i, so CP_i's mark-up is $m_i = r_i - c_i$ and the profit is $m_i \lambda \varphi_i$, where λ is the

[3]The reasons can be lack of time (for instance, a spanish user may watch news online through the TVE or the Antena3 website), or because their friends use mainly one chat (i.e. Messenger vs. Skype), among others.

[4]Each request can measured by each user's clicks on the internet page associated to the content, the λ parameter of our model.

demand intensity of requests and σ_i is the market share of CP_i.

Timing. The timing of the game is the following:

1. Each ISP_j decides its network capacity μ_j. This stage is only played in the long run. In the short run μ_j is fixed.

2. ISPs set subscription prices to users, p_j.

3. Each ISP negotiates the priority fee f_j with a CP. This stage is only played under network discrimination.

4. Users observe prices and capacities (and, under discimination, priorities) and subscribe to one ISP.

5. Given the ISP subscription, users browse one CP.

The model is solved by backward induction and the equilibrium concept is the sub-game perfect Nash equilibrium.

Market representation. The market we consider can be represented as a variation of a two-dimensional Hotelling model where users first choose one component (the online service) and then choose the other component (the content) of the final product (the online and content services) . In Figure 4..1 , we represent one example of how the market is shared among firms.

Since ISP_A provides service type $y = 0$, all users type $(x, 0)$, $\forall\ x \in [0, 1]$, do not incur in transport costs when subscribing to ISP_A. Similarly, ISP_B provides the

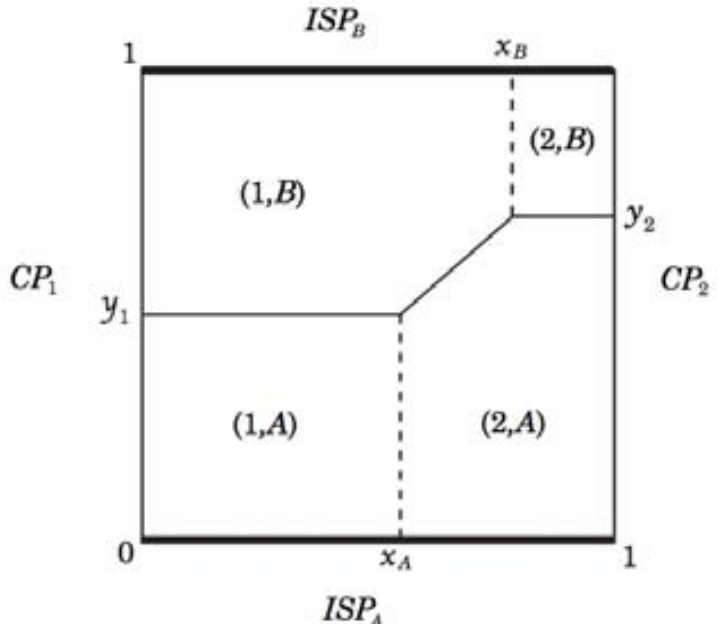

Figure 4..1: Market representation

preferred service to users type $(x, 1)$. CP_1 distributes content $x = 0$, which exactly meets the users type $(0, y)$ preference, $\forall\ y \in [0, 1]$. CP_2 distributes a content type $x = 1$ that is the most preferred of users type $(1, y)$.

Users type (x_A, y) are indifferent between CP_1 and CP_2 if they connect to ISP_A. Similarly, users (x_B, y) are indifferent between CPs when connecting to ISP_B. Users (x, y_1) are indifferent between ISP_A and ISP_B when browsing CP_1, while users (x, y_2) are indifferent when browsing CP_2.

In Figure 4..1, region $(1, A)$ represents the market share of users that browse CP_1 through ISP_A. Similarly, area $(2, A)$ represents the market share of users who browse CP_2 through ISP_A. The remaining areas refer to ISP_B. Hence, regions $(1, A)$ and $(2, A)$ represent ISP'_As market share, while regions $(1, A)$ and $(1, B)$ represent CP'_1s market share.

Network congestion To introduce network congestion in a competitive framework, we provide a new approach that is able to capture some of the qualitative features of queuing theory and that are interesting to our model. It simplifies the analysis and allows tractable solutions.

Under network neutrality, the speed at which contents are delivered to users is the same, so each user subscribing ISP_j has the expected waiting disutility of w, given by:

$$w = y - \left(\mu_j - \lambda\right)$$

where y is each user's perception of ISP_j's market share, λ is the demand intensity of content and μ_j is the ISP_j's network capacity, where $\mu_j > \lambda$. The disutility increases with the perceived market share of the ISP, i.e., with more people accessing the network. Also, it increases with the demand intensity and decreases with the ISP's network capacity. As already mentioned, users decide which ISP to subscribe based on the market share "perception" of that ISP. The rationale is that a user knows the share y of users of her type x that connect to her ISP, but she cannot know the share of other types x who subscribe the same ISP.[5]

Under net discrimination, the expected waiting disutility of a user requesting a priority content is w^p, given by:

$$w^p = w - \alpha(1-x)\left(\frac{1}{\mu_j} - \lambda\right) = y - \left(\mu_j - \lambda\right) - \alpha(1-x)\left(\frac{1}{\mu_j} - \lambda\right) \qquad (4..1)$$

[5]Literature in industrial organization provides many cases where the rationality hypothesis is dropped. See, for instance, the review of Ellison (2006). In our context, for instance, consumers are assumed to be able to solve a massive game theory problem with all other consumers. In the telecommunication framework, an interesting example of how the bounded rationality assumption better fits with empirical evidence, is provided in Mobius (2001).

where w^p is the difference between the waiting disutility w under net neutrality and an amount of time that depends on the share of users that request content from the priority class x, and on the degree of priority the ISP imposes on the content $\alpha > 0$. In contrast, the user that requests content without priority faces an expected waiting disutility of w^d:

$$w^d = w + \alpha x \left(\frac{1}{\mu_j} - \lambda \right) =$$

$$= y - \left(\mu_j - \lambda \right) + \alpha x \left(\frac{1}{\mu_j} - \lambda \right) \tag{4..2}$$

where here α reflects the extra disutility of requesting a content without priority.[6]

Assumption 1 ISPs' network capacity is always enough to serve users' requests, that is, $\lambda \mu_j < 1, j = A, B$.

As in Choi and Kim's model for internet congestion, our approach satisfies the three properties. First, each user faces a higher waiting disutility when requesting a second-priority content instead of a prioritized one, that is,

Property 1 $w^d > w > w^p$

This property is established by computing the difference between w^p and w^d. Also, the difference between waiting disutilities is constant regardless of the distribution of total traffic across different priority classes. Second, we find that the difference in waiting disutility becomes smaller as the network capacity increases, that is,

Property 2 $\frac{\partial (w^d - w^p)}{\partial \mu} < 0$

[6] If α was set to 0, the priority effect is null and we would be in the neutral regime.

The marginal reduction in waiting disutility for the priority service from an expansion in ISP capacity expansion decreases as the network capacity level becomes high. The intuition is that as an ISP increases network capacity, other things being equal, there is less congestion and priority becomes relatively less attractive compared to the non-priority service. Third, net discrimination does not change the share-weighted average waiting disutility:

Property 3 $xw^p + (1-x)w^d = w$

Assigning priorities does not change aggregate waiting disutility. It is only a way to manage content delivery. Some users improve their utility while others deteriorate due to higher waiting disutility, but the wheighted average waiting disutility of a network remains unchanged.

4.3 ISP Monopoly

Given that our model is in the spirit of Choi and Kim (2010), this section analyses the case where there is only one ISP and users have heterogeneous preferences with respect to contents. When using our approach to network congestion, the same qualitative results are attained as in Choi and Kim with respect to the long run incentives of ISPs to invest in network capacity, both under a neutral network and a discriminatory network regime.

Short run Analysis Under net neutrality, users pay a subscription price p to the ISP (with monopoly, $y = 1$) and choose one of two CPs. The indifferent user x^* between

the two contents is defined as

$$v - w - tx^* - p = v - w - t(1 - x^*) - p$$

where users $x \leq x^*$ browse CP_1 and users $x > x^*$ browse CP_2. Since CPs are located at the extremes of the preference line, $x^* = \frac{1}{2}$, i.e., the market is shared equally. The ISP sets the price p to maximize profit, π_M, conditional on full market coverage. This implies a positive utility for the indifferent user $x = \frac{1}{2}$. The equilibrium profit of the ISP is $\pi_M^* = p^* = v - 1 + \mu - \lambda - \frac{t}{2}$.

Under net discrimination, the monopolist can charge a fee f to the CP that purchases the priority, so the waiting time differs between contents. The user $\tilde{x}$ that is indifferent between the priority content and the non-priority content is

$$v - w^p(\tilde{x}) - t\tilde{x} - p = v - w^d(\tilde{x}) - t(1 - \tilde{x}) - p.$$

where the tilde is used to denote the variables under the discriminatory regime. The solution is $\tilde{x} = \frac{1}{2} + \alpha\frac{1-\mu\lambda}{2t\mu}$ and we easily observe that the priority CP has the largest market share, $\tilde{x} > \frac{1}{2}$. In an interior solution where both CPs operate in the market (i.e. a sufficiently high transport cost $t > \alpha\frac{1-\mu\lambda}{\mu}$), the market share of the priority CP is stable and decreases with the ISP's network capacity, i.e., $\frac{\partial\tilde{x}}{\partial\mu} < 0$. This result follows Choi and Kim. The ISP sets the subscription price to maximize profit, given by $\tilde{p} + f$, conditional on full market coverage. The priority fee f is calculated through Nash bargaining, and is $f = [m_2 + \theta(m_1 - m_2)](2\tilde{x} - 1)\lambda$.[7] Therefore, the ISP's profit in the discriminatory network is

$$\tilde{\pi}_M = \tilde{p} + f = (v - 1 + \mu - \lambda\tilde{x} - t\tilde{x}) + [m_2 + \theta(m_1 - m_2)](2\tilde{x} - 1)\lambda.$$

[7]The fee is an amount set between CP_2's and CP_1's willingness to pay for the priority service.

Long Run Analysis Now the network capacity in no longer fixed and the incentives to invest in expanding it are reflected in the partial derivatives of profits w.r.t. capacity. Under net neutrality, $\frac{\partial \pi_M^*}{\partial \mu} = 1$, so there is always incentive to invest. Under network discrimination, we obtain:

$$\frac{\partial \tilde{\pi}_m}{\partial \mu} = 1 - \frac{\alpha(1-\tilde{x})}{\mu^2} - \left(\frac{\alpha}{\mu} - \alpha\lambda + t\right)\frac{\partial \tilde{x}}{\partial \mu} + 2\lambda(m_1 + \theta(m_1 - m_2))\frac{\partial \tilde{x}}{\partial \mu}$$

To check whether the incentives to invest are higher under the discriminatory regime than under the neutral regime, we study the sign of the difference $(\frac{\partial \tilde{\pi}_M}{\partial \mu} - \frac{\partial \pi_M^*}{\partial \mu})$. As in Choi and Kim, the sign is undetermined. The effect of capacity expansion on the sale price of the priority is negative ($2\lambda(m_1 + \theta(m_1 - m_2))\frac{\partial \tilde{x}}{\partial \mu} < 0$), while the effect of capacity expansion on the subscription price due to discrimination is undetermined ($\frac{\alpha(1-\tilde{x})}{\mu^2} - (\frac{\alpha}{\mu} - \alpha\lambda + t)\frac{\partial \tilde{x}}{\partial \mu}$).

Hence, for solutions where both CPs serve the market, the overall effect of incentives to invest more under the discriminatory regime than under the neutral regime is undefined. On one hand, under net neutrality the monopolist always has incentives to invest in network capacity. On the other hand, under a discriminatory regime, the monopolist continues to face two effects that may go in the opposite direction. The effect of expanding capacity on the priority fee is negative, since less users choose the priority content, but the effect on the end-user subscription fee in undetermined. So, under our approach we show that there can be cases where the ISP may invest more under net discrimination than under net neutrality.

4.4 Benchmark: ISP Competition and Network Neutrality

In this section, we analyze the equilibrium outcome when two ISPs compete for users and are not allowed to discriminate between contents. This represents a benchmark to assess if allowing network discrimination leads to lower or higher incentives in capacity expansion investment.

Under net neutrality, the waiting disutility of any user that subscribes ISP_j, $j = A, B$, is the same regardless of the content browsed. We denote the waiting disutility of a user who browses CP_i under ISP_j as $w_{i,j}$ and we have

$$w_{1,j} = w_{2,j} = y - (\mu_j - \lambda).$$

The waiting time may not be the same in the whole internet, but it is so in each of the two networks. The user that subscribes ISP_A and is indifferent between contents is defined as:

$$v - y + (\mu_A - \lambda) - tx - sy - p_A = v - y + (\mu_A - \lambda) - t(1 - x) - sy - p_A$$

which is exactly the same solution as in the monopoly case, $x^* = \frac{1}{2}$. The indifferent user that subscribes ISP_B is the same.

When choosing which ISP to subscribe to, the indifferent user is characterized by:

$$v - y + (\mu_A - \lambda) - sy - p_A = v - (1 - y) + (\mu_B - \lambda) - s(1 - y) - p_B. \qquad (4..3)$$

The solution is:

$$y^*(p_A, p_B, \mu_A, \mu_B) = \frac{1}{2} + \frac{p_B - p_A}{2(1 + s)} + \frac{\mu_A - \mu_B}{2(1 + s)}. \qquad (4..4)$$

$ISP_A's$ market share of users who browse CP_1 is $\sigma^*_{1,A} \equiv x^* y^* = \frac{y^*}{2}$. Similarly, $ISP_A's$ market share of users who browse CP_2 is $\sigma^*_{2,A} \equiv (1 - x^*)y^* = \frac{y^*}{2}$. Hence, the total market share of ISP_A is $\sigma^*_A \equiv \sigma^*_{1,A} + \sigma^*_{2,A} = y^*$, which depends on the network capacity difference, as well as on the price differential. It depends positively on its own network capacity but negatively on price and transport cost, and the opposite with respect to the competitor decisions on capacity, price and transport cost. We represent the market shares, under net neutrality, in Figure 4..2 .

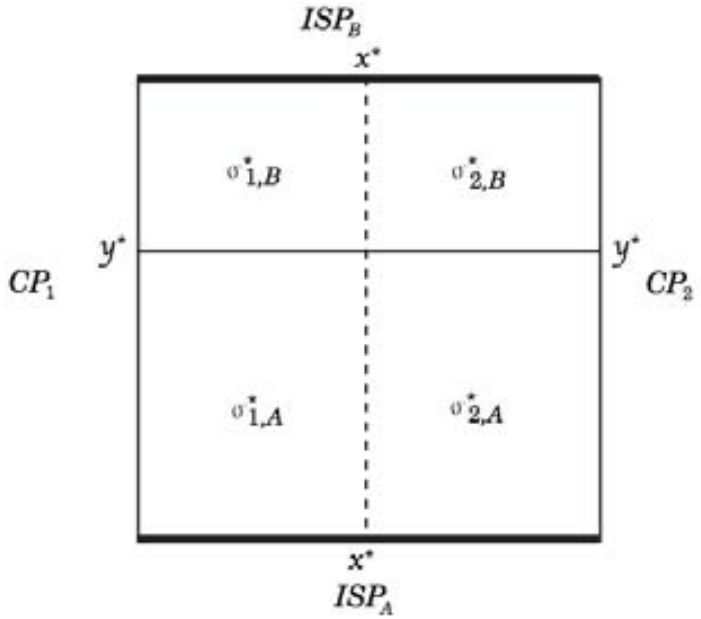

Figure 4..2: Market sharing under net neutrality

The equilibrium prices set by the ISPs is the Nash Equilibrium in ISP duopoly competition. For each ISP, the profit is the subscription price times its market share. The prices set by ISPs are the solution to:

$$\max_{p_A} \pi_A = y^*(p_A, p_B)\ p_A \qquad \text{s.t.} \quad v - y^* + (\mu_A - \lambda) - sy^* - p_A \geq 0$$

$$\max_{p} \pi_B = (1 - y^*(p_A, p_B))\, p_B \qquad \text{s.t.} \quad v - (1 - y^*) + (\mu_B - \lambda) - s(1 - y^*) - p_B \geq 0$$

and are the following:

$$p_A^* = 1 + s + \frac{\mu_A - \mu_B}{3}$$
$$p_B^* = 1 + s - \frac{\mu_A - \mu_B}{3}.$$

In order to guarantee that both ISPs operate in the market, we assume they are sufficiently differentiated relative to the network capacity differential:

Assumption 2 $s > \frac{(\mu_A - \mu_B)}{3} - 1$

Hence the equilibrium market share is:

$$\sigma_A^* = \frac{1}{2} + \frac{\mu_A - \mu_B}{6(1+s)} \tag{4..5}$$

and $ISPs'$ profits are:

$$\pi_A^* = \frac{(1+s)}{2} + \frac{(\mu_A - \mu_B)}{3} + \frac{(\mu_A - \mu_B)^2}{18(1+s)}$$

and

$$\pi_B^* = \frac{(1+s)}{2} - \frac{(\mu_A - \mu_B)}{3} + \frac{(\mu_A - \mu_B)^2}{18(1+s)}.$$

In the long run, ISPs always have incentives to invest in network capacity. ISP_A always has incentive to invest in network capacity, since $\frac{\partial \pi_A^*}{\partial \mu_A} > 0$. This is an intuitive result, since more relative capacity always provides a higher market share of end-users. For ISP_B, there is always incentive to increase network capacity since $\frac{\partial \pi_B^*}{\partial \mu_B} > 0$ under assumption 4.4.

4.5 ISP Competition and Network Discrimination

We now introduce the possibility of ISP competition under discrimination of contents, that is, ISPs are able to offer a tiered service. We compare the equilibrium outcomes with the ones of network neutrality in Section 4.6. We first consider the last stages where users choose the ISP they want to subscribe to and which content they want to browse. In the first stages, we analyze how ISPs assign the priority service to one content provider by charging a fee payment and setting subscription prices to users.

Users' choice After observing the announced priority service by the ISPs, the subscription prices, and the network capacity of both ISPs, users: (1) subscribe their preferred ISP and (2) decide which content to browse. The utility function of user type (x, y) when she browses CP_i through $ISPj$, is denoted as $u_{i,j}$ and is the following:

$$u_{i,j} = \begin{cases} v - w_{1,A}(\cdot) - xt - ys - p_A, & \text{if her choice is } (1, A) \\ v - w_{2,A}(\cdot) - (1-x)t - ys - p_A, & \text{if her choice is } (2, A) \\ v - w_{1,B}(\cdot) - xt - (1-y)s - p_B, & \text{if her choice is } (1, B) \\ v - w_{2,B}(\cdot) - (1-x)t - (1-y)s - p_B, & \text{if her choice is } (2, B) \end{cases}$$

where waiting disutilities depend on the content that ISPs prioritize.

2nd stage: CP choice Once users subscribe one ISP, they decide which content to browse. Among all users (x, y) with a given preference y for an ISP, a user who subscribes ISP_j is indifferent between CP_1 and CP_2 when $u_{1,j} = u_{2,j}$. For example, a user

who subscribes ISP_A and CP_1 is the priority content, then is has the following utilities:

$$u_{i,A} = \begin{cases} v - \left(y - (\mu_A - \lambda) - \alpha(1-x)\left(\frac{1}{\mu_A} - \lambda\right)\right) - xt - ys - p_A, & if\ i = 1 \\ v - \left(y - (\mu_A - \lambda) + \alpha x\left(\frac{1}{\mu_A} - \lambda\right)\right) - (1-x)t - ys - p_A, & if\ i = 2 \end{cases}$$

where waiting disutilities are defined according to equations (4..1) and (4..2). As subscribers of ISP_A, the users type $(\tilde{x}_A, y)$ that are indifferent between CP_1 and CP_2 are given by:

$$\tilde{x}_A = \frac{1}{2} + \alpha\frac{1 - \lambda\mu_A}{2t\mu_A}.$$

From the solution attained, clearly some users switch from the discriminated content to the priority content, as $\tilde{x}_A > \frac{1}{2}$. The second component captures the deviation from the half-half solution attained in net neutrality. We name it the *priority effect*: a fraction of users $\alpha\frac{1-\lambda\mu_A}{2t\mu_A}$ switches to CP_1 since it provides lower waiting disutility. Since, from the users' perspective, CPs are symmetric, then if ISP_A prioritizes CP_2 the indifferent user is $(1 - \tilde{x}_A, y)$, where

$$1 - \tilde{x}_A = \frac{1}{2} - \alpha\frac{1 - \lambda\mu_A}{2t\mu_A}.$$

To avoid excessive notation, we denote as $(1 - \tilde{x}_A, y)$ and $(1 - \tilde{x}_B, y)$ the indifferent users subscribers of ISP_A and ISP_B, respectively, when CP_2 is the priority content.[8] Since networks are assymetric with respect to capacity, and hence, delivery speed of contents, this priority effect is different between ISPs. Proposition 4..1 shows the effect of congestion externality between networks on the share of users between content providers.

[8] We look for an interior equilibrium where the users browse both CP_1 and CP_2 in both ISPs. One necessary condition is $t > \alpha\frac{1-\lambda\mu_B}{\mu_B}$, implying that CPs are sufficiently differentiated so there is always an indifferent x-type user between them.

Proposition 4..1 *The priority content in the small network attracts a larger proportion of users of a given* y*-type than in the large network, that is,*

$$1 - \tilde{x}_B < 1 - \tilde{x}_A < x^* < \tilde{x}_A < \tilde{x}_B.$$

Since ISP_A is the large network, contents are delivered faster and its users face lower disutility than ISP_B users. Therefore, the congestion effect is stronger in ISP_B and its users switch more often from their preferred CP to the other CP than ISP_A users. When ISP_A sells the priority service to CP_1, its users' waiting disutilities are:[9]

$$\begin{aligned} w_{1,A}(\tilde{x}_A) =& y - \mu_A + \lambda - \alpha(1 - \tilde{x}_A)\left(\frac{1}{\mu_A} - \lambda\right) \\ w_{2,A}(\tilde{x}_A) =& y - \mu_A + \lambda + \alpha\tilde{x}_A\left(\frac{1}{\mu_A} - \lambda\right). \end{aligned}$$

Similar disutilities are associated with ISP_B (when a user connects to ISP_B, waiting disutilities depend on its network capacity, μ_B, and on the perceived congestion, $1-y$). When one ISP assigns priority to one CP, the waiting disutility of the user that browses it decreases, while the disutility of the user browsing the discriminated content increases. Moreover, as expressed in proposition 4..2, the waiting disutility gaps of users do not change, independently of the prioritized content.

Proposition 4..2 *Under net discrimination,the waiting disutility gap difference of a user browsing the priority CP in the small network and in the large network is inde-*

[9] If it sells the priority service to CP_2, the waiting disutilities are

$$\begin{aligned} w_{1,A}(1 - \tilde{x}_A) =& y - \mu_A + \lambda + \alpha(1 - \tilde{x}_A)\left(\frac{1}{\mu_A} - \lambda\right) \\ w_{2,A}(1 - \tilde{x}_A) =& y - \mu_A + \lambda - \alpha\tilde{x}_A\left(\frac{1}{\mu_A} - \lambda\right). \end{aligned}$$

pendent of which CP it prioritizes and is:

$$\omega(\mu_A, \mu_B) \equiv \frac{\alpha^2(\mu_A - \mu_B)(\mu_A + \mu_B - 2\lambda\mu_A\mu_B)}{2t\mu_A^2\mu_B^2}. \tag{4..6}$$

The gap $\omega(\mu_A, \mu_B)$ is positive since $\mu_A + \mu_B - 2\lambda\mu_A\mu_B > 0$. The results of the proposition are implied by Property 1 regarding the waiting disutility w, following the same intuition as in Choi and Kim. When ISPs prioritize one content, this creates a disutility gap between users browsing priority content and users browsing discriminated content. However, the gap is larger in smaller networks because, as stated by Choi andKim, larger capacity makes congestion less important. Hence, less users switch from the non-priority to the priority content and the average gap is lower (Property 2).

This intuition also applies when ISPs prioritize different CPs: within an ISP, the switching effect of users is the same, independently of the content that has been prioritized (since contents are symmetric to the users). The total waiting disutility faced by its users does not change when an ISP discriminates, since a fraction of users switch from the discriminated content to the prioritized CP, from Property 3.

1st stage: ISP choice Depending on the choice of CP, users anticipate the waiting disutilities according to their y-type. Hence, a user (x, y) receives utility

$$u_{i,j} = \begin{cases} v - w_{1,A}(\tilde{x}) - xt - ys - p_A, & \text{if her choice is } (1, A) \\ v - w_{2,A}(\tilde{x}) - (1-x)t - ys - p_A, & \text{if her choice is } (2, A) \\ v - w_{1,B}(\tilde{x}) - xt - (1-y)s - p_B, & \text{if her choice is } (1, B) \\ v - w_{2,B}(\tilde{x}) - (1-x)t - (1-y)s - p_B, & \text{if her choice is } (2, B) \end{cases} \tag{4..7}$$

where by $\tilde{x}$ we mean the prioritized content under an ISP. Using the example above where both ISPs sell the priority to CP_1, we have users:

- type $(x \leq \tilde{x}_A, y)$ that always browse CP_1. They choose which ISP to subscribe to according to their preference y and there is one indifferent type $(x \leq \tilde{x}_A, y = \tilde{y}_1)$ between ISPs (that is, $u_{1,A} = u_{1,B}$);

- users type $(x > \tilde{x}_B, y)$ always browse CP_2. They choose which ISP to subscribe to according to their preference y and there is one type $(x > \tilde{x}_B, y = \tilde{y}_2)$ that is indifferent between ISPs;

- users type $(x \in (\tilde{x}_A, \tilde{x}_B], y)$ browse CP_2 in ISP_A and CP_1 in ISP_B and there is one type $(x \in (\tilde{x}_A, \tilde{x}_B], y = \tilde{y}_m(x))$ that is indifferent between browsing CP_2 in ISP_A or CP_1 in ISP_B.

We represent the indifferent users of this case in Figure 4..3. For example, user $(\tilde{x}_A, \tilde{y}_1)$ receives the same utility if connecting to ISP_B and browsing CP_1 or if connecting to ISP_A and browsing either CP_1 or CP_2.

ISPs sell priority to the same CP Let's assume both ISPs sell the priority to CP_1. Then, users type $(x \leq \tilde{x}_A, y)$ who always choose CP_1, independently of the ISP, are indifferent for $(x \leq \tilde{x}_A, y = \tilde{y}_1)$ when $u_{1,A} = u_{1,B}$, that is[10]:

$$\tilde{y}_1 \equiv y^* + \frac{\omega(\mu_A, \mu_B)}{2(1+s)} - \frac{\alpha(\mu_A - \mu_B)}{4(1+s)\mu_A\mu_B}.$$

Here, y^* is the market share of ISP_A under net neutrality for these users. The

[10] The case where both ISPs sell the priority to CP_2 is symmetric to this case. Indifferent users type $(x \leq \tilde{x}_A, y = \tilde{y}_1)$ solve:

$$v - w_{1,A}(\tilde{x}_A, y) - xt - ys - p_A = v - w_{1,B}(\tilde{x}_B, 1-y) - xt - (1-y)s - p_B.$$

Note we added y in the waiting disutility: now y-type's perception of the congestion in the two networks affects her ISP choice.

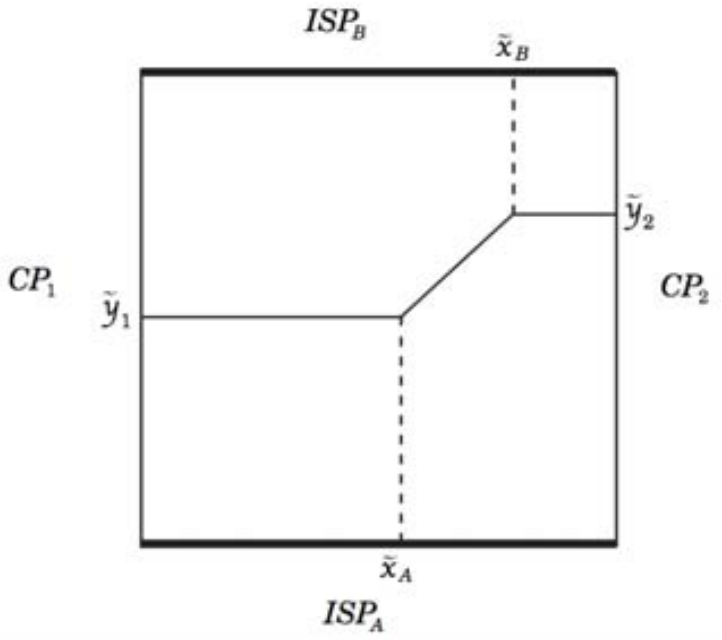

Figure 4..3: Indifferent users by type

term $\frac{\omega(\mu_A,\mu_B)}{2(1+s)}$ reflects the disutility gap difference between the users of two networks and the term $\frac{\alpha(\mu_A-\mu_B)}{4(1+s)\mu_A\mu_B} > 0$ denotes how much is the gap difference in the two networks due to comparing users across networks that browse non-priority content and that browse priority content. Also, users type $(x > \tilde{x}_B, y)$ who always choose CP_2, independently of the ISP, are indifferent between the two ISPs when $u_{2,A} = u_{2,B}$. In this case, these users are $(x > \tilde{x}_B, y = \tilde{y}_2)$, where[11]

$$\tilde{y}_2 = y^* + \frac{\omega(\mu_A, \mu_B)}{2(1+s)} + \frac{\alpha(\mu_A - \mu_B)}{4(1+s)\mu_A\mu_B}.$$

It is clear to see that the share of loyal users attracted by the larger network ISP_A is larger among the users of the non-prioritized content than of the prioritized

[11] Indifferent users type $(x \leq \tilde{x}_A, y = \tilde{y}_2)$ solve the following equation:

$$v - w_{1,A}(1 - \tilde{x}_A, y) - (1-x)t - ys - p_A = v - w_{1,B}(1 - \tilde{x}_B, 1-y) - (1-x)t - (1-y)s - p_B.$$

content, that is $\tilde{y}_1 < \tilde{y}_2$.

ISPs sell priority to different CPs Consider that ISP_A prioritizes CP_2 and ISP_B prioritizes CP_1.[12] Users type $(x \leq 1 - \tilde{x}_A, y)$ always browse CP_1, users type $(x > \tilde{x}_B, y)$ always browse CP_2, and users type $(x \in (1 - \tilde{x}_A, \tilde{x}_B], y)$ browse CP_2 under A and CP_1 under B.

Among the users type $(x \leq 1 - \tilde{x}_A, y)$ that always choose CP_1 independently of the ISP they join to, the indifferent users are[13]

$$\hat{y}_1 = y^*(p_A, p_B; \mu_A, \mu_B) + \frac{\omega(\mu_A, \mu_B)}{2(1+s)} - \frac{\alpha(\mu_A + \mu_B - 2\lambda\mu_A\mu_B)}{4(1+s)\mu_A\mu_B}$$

where we denote by $\hat{y}$ the fact that indifferent users are determined by priorities assigned to different CPs. Also, among users type $(x > \tilde{x}_B, y)$ that always choose CP_2 independently of the ISP they have subscribed, the indifferent users are[14]

$$\hat{y}_2 = y^*(p_A, p_B; \mu_A, \mu_B) + \frac{\omega(\mu_A, \mu_B)}{2(1+s)} + \frac{\alpha(\mu_A + \mu_B - 2\lambda\mu_A\mu_B)}{4(1+s)\mu_A\mu_B}$$

Now, compared to the case where both ISPs sell the priority to CP_1, ISP_A attracts more users who browse CP_2: it prioritizes a content that is discriminated in ISP_B. Hence, the waiting disutility difference between A and B is emphasized. Similarly, ISP_B attracts more users who browse CP_1. In Figure 4..4, we represent the indifferent

[12] The other case, where CP_1 is prioritized by A and CP_2 by B, is symmetric.

[13] Indifferent users satisfy the following condition:

$$v - w_{1,A}(1 - \tilde{x}_A, y) - xt - ys - p_A = v - w_{1,B}(\tilde{x}_B, 1 - y) - xt - (1-y)s - p_B.$$

[14] Indifferent users satisfy the following condition:

$$v - w_{1,A}(\tilde{x}_A, y) - (1-x)t - ys - p_A = v - w_{1,B}(1 - \tilde{x}_B, 1 - y) - (1-x)t - (1-y)s - p_B.$$

users in this case.

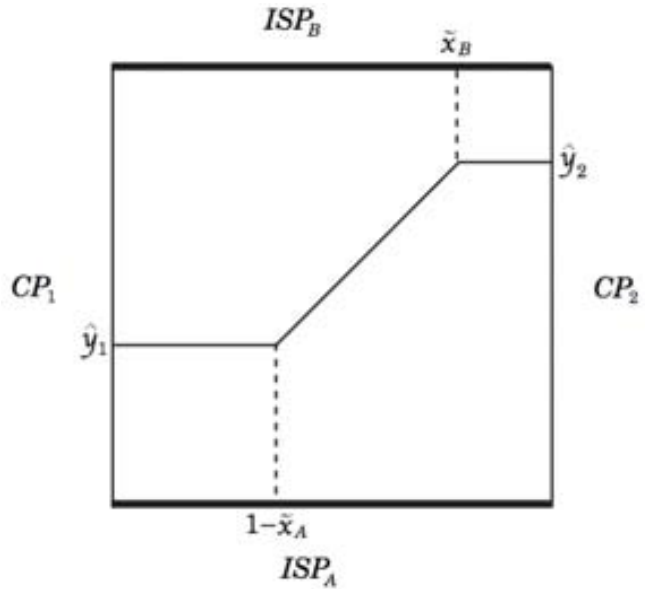

Figure 4..4: Indifferent users when priorities are sold to different CPs

Proposition 4..3 *A discriminating network attracts more loyal users of the content it prioritizes when the other network prioritizes a different content. That is,*

$$\hat{y}_1 < \tilde{y}_1 < \tilde{y}_2 < \hat{y}_2$$

Since ISP_A is the large network, its users face a lower congestion disutility. When contents are treated differently, then less users switch to the premium content in comparison to ISP_B since they are less penalized by browsing a content that is not their most preferred.

Ceteris paribus, ISP_B is less attractive than in case of network neutrality. The lower attractiveness of B is the same, regardless if ISPs prioritize the same content or if ISPs prioritize different CPs. However, both ISPs attract more users of the priority

content if they sell the priority to different CPs. This effect is, of course, compensated by a higher loss of users that browse the discriminated content. Overall we cannot readily say that the market share of ISP_A is higher, but we infer that there is less ISP competition to attract users that browse the same content independently of the ISP, since there is a higher product differentiation in this case.

Market shares The market share of each ISP will depend on which CP it prioritizes. Four different priority combinations can occur. Given that CPs are symmetric from users' perspective, each ISP obtains the same market share when both prioritize the same CP, regardless which content it is. Similarly, an ISP's market share does not change if it prioritizes CP_1 and the competitor prioritizes CP_2 or in the opposite case. Therefore, we study two cases: both ISPs sell the priority to CP_1; ISP_A prioritizes CP_2 and ISP_B prioritizes CP_1. The two other cases yield the same results as the two cases we explore now.

Both ISPs sell the priority to CP_1 The market share of ISP_A is given by the users who browse CP_1 (denoted by $\tilde{\sigma}_{1,A}$) and CP_2 (denoted by $\tilde{\sigma}_{2,A}$) in this network.[15] So, the share of users connected to A are $\tilde{\sigma}_A \equiv \tilde{\sigma}_{1,A} + \tilde{\sigma}_{2,A}$:

$$\tilde{\sigma}_A = \tilde{x}_A\tilde{y}_1 + (1-\tilde{x}_B)\tilde{y}_2 + \frac{(\tilde{y}_1+\tilde{y}_2)(\tilde{x}_B-\tilde{x}_A)}{2}.$$

[15] In particular, the total amount of users who always browse CP_1 is $\tilde{\sigma}_{1,A} = \tilde{x}_A \cdot \tilde{y}_1$ and the size of users who browse CP_2 is $\tilde{\sigma}_{2,A} = (1-\tilde{x}_B)\cdot\tilde{y}_2 + \frac{(\tilde{y}_1+\tilde{y}_2)(\tilde{x}_B-\tilde{x}_A)}{2}$, where the last fraction represents the share of users type $(x \in (\tilde{x}_A, \tilde{x}_B], y)$ who browse CP_2 when they join to ISP_A.

Similarly, $ISP_B's$ market share is

$$\tilde{\sigma}_B = \tilde{x}_A(1-\tilde{y}_1) + (1-\tilde{x}_B)(1-\tilde{y}_2) + \frac{(2-(\tilde{y}_1+\tilde{y}_2))(\tilde{x}_B-\tilde{x}_A)}{2}.^{16}$$

ISP_A sells priority to CP_2 and ISP_B to CP_1 In this case, the market share of ISP_A is composed of users who browse CP_1 (denoted as $\hat{\sigma}_{1,A}$) and CP_2 ($\hat{\sigma}_{2,A}$) when A prioritizes CP_2 and B prioritizes CP_1: $\hat{\sigma}_A \equiv \hat{\sigma}_{1,A} + \hat{\sigma}_{2,A}$. Hence, the total market share is:

$$\hat{\sigma}_A = (1-\tilde{x}_A)\hat{y}_1 + (1-\tilde{x}_B)\hat{y}_2 + \frac{(\hat{y}_1+\hat{y}_2)(\tilde{x}_B-(1-\tilde{x}_A))}{2}.$$

Also, the market share of ISP_B is

$$\hat{\sigma}_B = (1-\tilde{x}_A)(1-\hat{y}_1) + (1-\tilde{x}_B)(1-\hat{y}_2) + \frac{(2-(\hat{y}_1+\hat{y}_2))(\tilde{x}_B-(1-\tilde{x}_A))}{2}.$$

In Figure 4..5, we provide an illustration of market shares. When both ISPs prioritize the same content, ISP_A serves a larger proportion (compared to ISP_B) of users browsing the discriminated content, as Figure 4..5 (a) shows. This is the intuition we withdraw from the fact that congestion is less valuable in the larger network and users browsing the discriminated content are less penalized. When ISPs prioritize different contents, there is an increase in the proportion of users browsing their respective premium contents, as shown in Figure 4..5 (b).

Market shares comparison Before assigning priorities, ISPs evaluate the impact on their market shares, i.e., whether prioritizing the same content as the other ISP attracts more users than prioritizing the other content. Interestingly, market share is not affected by priority.

[16] It is worthwhile remembering that users' content choice is a function of network capacities, that is $\tilde{x}_i(\mu_A, \mu_B)$. Also, the subscription service choice is $\tilde{y}_i(p_A, p_B; \mu_A, \mu_B)$.

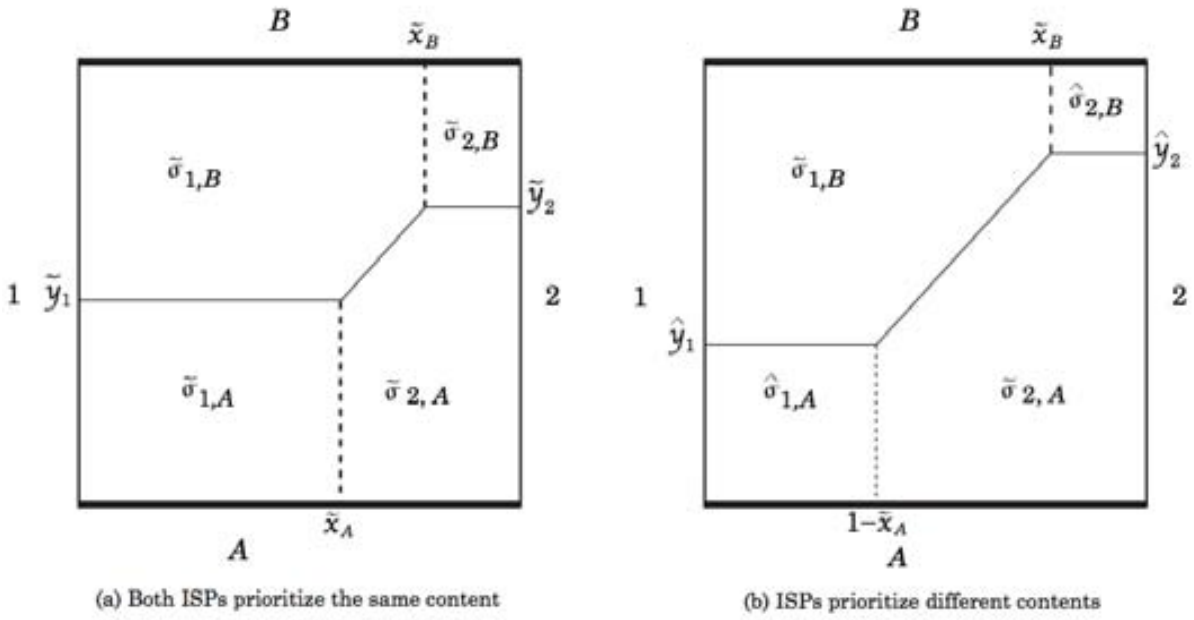

Figure 4..5: Representation of ISP market shares

Proposition 4..4 *Under network discrimination, each ISP's market share is the same independently to which CP the priority is sold. In particular, $ISP'_A s$ market share increases by the waiting disutility gap $\frac{\omega}{4(1+s)}$ compared with the case of network neutrality. Symmetrically, $ISP'_B s$ market share decreases by the waiting disutility gap $\frac{\omega}{4(1+s)}$.*

The fact that total market shares do not depend on priority assignments stems from Property 3. The total disutility does not change when an ISP prioritizes a content: the disutility gap is inversely proportional to the number of users who face lower quality. This implies that the number of users who would migrate to the larger network is constant.

This result extends one of the results of Choi and Kim. In their model, when a monopolist ISP discriminates contents, the utility of users browsing the discriminated

content deteriorates. Therefore, the monopolist must set a lower subscription price to guarantee full market coverage when participation is inelastic. In this model, market participation is also inelastic but ISPs' shares are completely elastic. The users that browse discriminated content see their utility deteriorate more in one ISP than in the other. Observinf that, they can switch ISP. In particular, the users in the small network switch to the large network. It is up to the small network to either reduce the subscription price to maintain users or keep the price but lose those users.

Priority pricing We have seen before that ISPs' priority strategies do not affect market share, hence each ISP sells the priority to the content that provides the highest revenue. On the other hand, a CP purchases the priority service if the extra-profits are greater or equal to the priority fee. The $CP_i's$ willingness to pay is the improvement of its market share, denoted by $\Delta\tilde{\sigma}_i$, times per-user profit: $\lambda m_i \Delta\widetilde{\sigma}_i$.

If ISP_B assigns the priority service to CP_1, ISP_A can either assign the priority to CP_1 or sell it to CP_2. Both situations are represented in Figure 4..6.

In Figure 4..6 (a), ISP_A sells the priority service to CP_1; dotted lines represent market shares in case the priority be sold to CP_2. The market share gained by CP_1 with respect to when ISP_A sells the priority to CP_2, is given by the hatched area. The horizontal lines represent the users of ISP_A who shift from CP_2 to CP_1 given the higher speed of the content provider. Vertical lines denote the users of ISP_A who were browsing CP_2 but, given the worsened speed, have shifted to ISP_B choosing CP_1. The shaded area represents the users of ISP_B browsing CP_1 and who have now shifted to ISP_A because it is their preferred ISP; these users do not constitute an effective change

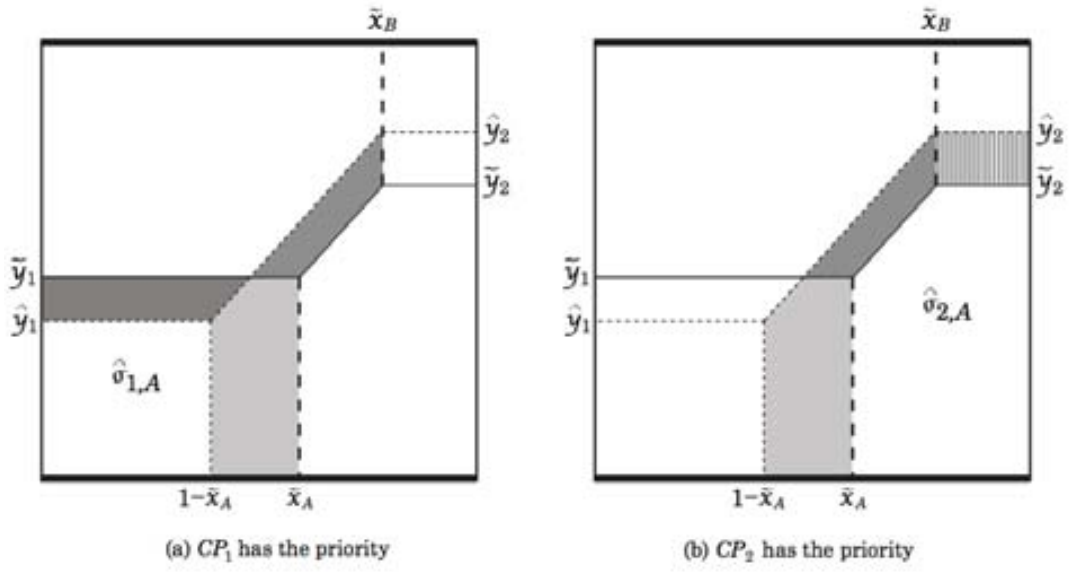

Figure 4..6: Users gained by a CP with priority under ISP_A

in CP_1's market share. In Figure 4..6 (b), ISP_A assigns the priority service to CP_2. The users who shift their choice are similar to the previous case. If ISP_B sells the priority to CP_2, the result is symmetric.

When ISP_B decides which CP to sell the priority to, the problem is similar to that one of ISP_A. We describe the problem in detail in Appendix 4.8.

Proposition 4..5 *The share of users gained by a CP that purchases the priority service from an ISP is always the same independently of which CP it is and of the priority decision of the other ISP. The improvement in CPs' market shares in the two networks are:*

$$\Delta\tilde{\sigma}_i = \frac{\alpha(1-\lambda\mu_A)}{t\mu_A}\left(y^* + \frac{\omega}{2(1+s)}\right), i = 1,2 \qquad \text{, under } ISP_A$$

$$\Delta\tilde{\sigma}_i = \frac{\alpha(1-\lambda\mu_B)}{t\mu_B}\left(1 - y^* - \frac{\omega}{2(1+s)}\right), i = 1,2 \qquad \text{, under } ISP_B$$

Note that the market share gained by CP_1 is proportional to the respective market size of each ISP (the terms in brackets). Hence, ISP_A can provide a higher share, due to the larger capacity of its network. However, more users switch from CP_2 to CP_1 in network B because of the prioritized content is more valuable when capacity is low (the fraction before brackets). The overall effect depends on the end-users price decisions of the ISPs.

Given that CP_1's margin is higher than CP_2's, ISP_A always prefers to sell the priority right to CP_1. Priority fee results from the Nash bargaining between the ISP and the CP: $(\theta m_1 + (1-\theta)m_2)\lambda\Delta\widetilde{\sigma}_1$. So, ISP_A charges a fee:

$$f_A = \frac{\alpha(1-\lambda\mu_A)}{t\mu_A}\left(y^* + \frac{\omega}{2(1+s)}\right)\lambda(\theta m_1 + (1-\theta)m_2).$$

Similarly, ISP_B also sells the priority right to CP_1. Hence the fee is $(\theta m_1 + (1-\theta)m_2)\Delta\widetilde{\sigma}_1$:

$$f_B = \frac{\alpha(1-\lambda\mu_B)}{t\mu_B}\left(1 - y^* - \frac{\omega}{2(1+s)}\right)\lambda(\theta m_1 + (1-\theta)m_2).$$

End-users pricing In case of network discrimination, ISP's profits are given by the price paid by the respective end-users plus the fee charged to the prioritiy content:

$$\pi_A = p_A\tilde{\sigma}_A + f_A$$

$$\pi_B = p_B\tilde{\sigma}_B + f_B$$

Both ISPs maximize their profits by choosing their respective access prices, p_j, where the only restriction is represented by full market coverage.[17] We assume the transport

[17] Therefore, the indifferent consumer is always receiving a nonnegative surplus when joining to one ISP. However we assume v, the parameter denoting the benefit from joining a network and browsing a content, is sufficiently large to prevent the participation constraint to bind.

cost s is sufficiently large to insure interior solutions.[18]

Proposition 4..6 *The asymmetric ISPs market has a unique interior equilibrium characterized by the following prices:*

$$\tilde{p}_A = p_A^* + \frac{\omega}{6} - \frac{\alpha(\mu_A + 2\mu_B - 3\lambda\mu_A\mu_B)}{3t\mu_A\mu_B}\lambda(m_2 - \theta(m_2 - m_1))$$
$$\tilde{p}_B = p_B^* - \frac{\omega}{6} - \frac{\alpha(2\mu_A + \mu_B - 3\lambda\mu_A\mu_B)}{3t\mu_A\mu_B}\lambda(m_2 - \theta(m_2 - m_1)).$$

Under discrimination, the equilibrium price differential increases with respect to the network neutrality case :

$$\tilde{p}_A - \tilde{p}_B = p_A^* - p_B^* + \frac{1}{3}\left(\omega + \frac{\alpha(\mu_A - \mu_B)}{t\mu_A\mu_B}\lambda(\theta m_1 + (1-\theta)m_2)\right).$$

When networks are allowed to discriminate between contents, a proportion of the fees charged to content providers is deducted in the price paid by users. This discount stems from higher ISP competition for users: a larger market share represents a higher revenue from the fee charged to the priority CP. Moreover, users who browse the discriminated content are more penalized in the small network (because of the higher waiting disutility). Hence, the small ISP needs to allow a higher discount (here denoted by ω) than the large ISP.

[18] There must be an indifferent user between ISP_A and ISP_B who browses CP_1 and an indifferent user who browses CP_2. This requires $\tilde{y}_1, \tilde{y}_2, \hat{y}_1, \hat{y}_2 \in (0,1)$, which depends on the equilibrium prices. Given that $\mu_A > \mu_B$ the condition to be satisfied is $\tilde{y}_2 < 1$. Therefore, the transport cost s must satisfy

$$s > \frac{\mu_A - \mu_B}{3} + \frac{2}{3}\omega - \frac{\alpha(\mu_A - \mu_B)\lambda(\theta m_1 + (1-\theta)m_2)}{3t\mu_A\mu_B} + \frac{\alpha(\mu_A + \mu_B - 2\lambda\mu_A\mu_B)}{2\mu_A\mu_B} - 1.$$

For simplicity, we denote by

$$\varphi \equiv \frac{\alpha(\mu_A - \mu_B)}{t\mu_A\mu_B}\lambda(\theta m_1 + (1-\theta)m_2) > 0 \tag{4..8}$$

the extra-discount that ISP_B allows as a proportion to the fee charged to the content provider; that is, a higher share of CP's margin is transferred to the users in the smaller network. The other term in the difference between equilibrium prices, ω, represents the extra-discount that ISP_B needs to allow whatever the margin of the CP is. In other words, both ISPs compete allowing higher discounts to users, in order to gain a larger market share to offer to the prioritized CP. In equilibrium, the smaller network offers a higher discount by φ. If the content provider has a high margin, ISPs compete for users aggressively and the discount to compensate extra-disutility in the small network (ω) is less relevant to the equilibrium prices. Otherwise, if the margin of the prioritized CP is insignificant, ISPs do not compete so fiercely through discounts. Therefore, the smaller network needs only to compensate the extra-utility loss ω.

Assumption 3 $\varphi > \omega$.

That is, the discount that ISP_B gives to its users to increase the market share is always high enough to compensate discriminated users for the lower utility. This requires:

$$\lambda(\theta m_1 + (1-\theta)m_2) > \frac{\alpha(1-\lambda\mu_A)}{2\mu_A} + \frac{\alpha(1-\lambda\mu_B)}{2\mu_B}$$

This assumption is related to how profitable prioritizing content is. If some users are penalized, then under this assumption ISPs always find to profitable to compensate

them with a price discount since the fee paid by the priority CP content is sufficiently large.

Proposition 4..7 *Under network discrimination, the equilibrium market share of the small network increases compared to the case of network neutrality. In particular, ISPs' equilibrium market shares are:*

$$\tilde{\sigma}_A = \sigma_A^* - \frac{1}{6(1+s)}\left(\varphi - \frac{\omega}{2}\right)$$

$$\tilde{\sigma}_B = \sigma_B^* + \frac{1}{6(1+s)}\left(\varphi - \frac{\omega}{2}\right)$$

ISP_B increases its market share since, in its network, priority is more valuable. More users switch to priority content, allowing ISP_B to charge a higher fee. ISP_A also has the incentive to offer a discount to attract more users, but since the users that browse the discriminated content are the less penalized, therefore, they are less eager to switch content and so there is less incentive to do so. Note that for an interior solution, $ISP_A's$ market share is still larger than $ISP_B's$. Finally, equilibrium fees charged to the priority CP are:

$$\tilde{f}_A = \frac{\mu_B(1-\lambda\mu_A)}{\mu_A - \mu_B}\left(\sigma_A^* - \frac{1}{3(1+s)}\left(\frac{\varphi}{2} - \omega\right)\right)\varphi$$

$$\tilde{f}_B = \frac{\mu_A(1-\lambda\mu_B)}{\mu_A - \mu_B}\left(\sigma_B^* + \frac{1}{3(1+s)}\left(\frac{\varphi}{2} - \omega\right)\right)\varphi.$$

$$\tilde{f}_A = \frac{\alpha(1-\lambda\mu_A)}{t\mu_A}\left(\sigma_A^* - \frac{1}{3(1+s)}\left(\frac{\varphi}{2} - \omega\right)\right)\lambda(\theta m_1 + (1-\theta)m_2)$$

$$\tilde{f}_B = \frac{\alpha(1-\lambda\mu_B)}{t\mu_B}\left(\sigma_B^* + \frac{1}{3(1+s)}\left(\frac{\varphi}{2} - \omega\right)\right)\lambda(\theta m_1 + (1-\theta)m_2)$$

Again, for an interior solution the fees are strictly positive.

4.6 Investment incentives

In the long run, ISPs can invest to increase network capacity μ_j. As in case of a monopolistic service provider, expanding capacity has two opposite effects on an ISP's revenues. On the one hand, higher capacity reduces the congestion in the network. Users achieve higher utility when connected (due to the reduced waiting disutility) and, therefore an ISP can charge a higher price to provide network connection. On the other hand, capacity expansion affects revenues from priority: lower congestion reduces the value for priority since users are less likely to switch to a prioritized content. Hence, a CP will pay a lower fee to purchase the premium service.

There are also two countervailing effects that only arise with ISP competition. First, if a network expands its capacity more than its competitor, it becomes more attractive to marginal users, given that they face lower congestion (and disutility) if they migrate to the network that invests more in capacity. Hence, the increased market share allows an ISP to gain higher revenues from end-users. Second, the lower fee charged to the CP reduces the discount an ISP can allow to its users. Therefore, the competitor can contrast users migration by offering a greater discount.

Equilibrium profits To analyze the overall effect of network discrimination on investment incentives, equilibrium profits of the two networks are compared to the case where network discrimination is not allowed (network neutrality). When an ISP plans investments in capacity expansion, it decides the optimal level of μ_j given the impact of capacity on equilibrium prices, fees, and market shares analyzed in the previous section.

In particular, for any capacity level μ_j, equilibrium profits of ISP_j are

$$\tilde{\pi}_j = \pi_j^* + \Gamma_j(\mu_A, \mu_B) \tag{4..9}$$

where for ISP_A $\Gamma_A(\mu_A, \mu_B)$ is defined as

$$\begin{aligned}\Gamma_A(\mu_A, \mu_B) \equiv & \frac{1}{18(1+s)}\left(\varphi - \frac{\omega}{2}\right)^2 - \left(\frac{1}{3} + \frac{\mu_A - \mu_B}{9(1+s)}\right)\left(\varphi - \frac{\omega}{2}\right) + \\ & + \frac{(1-\lambda\mu_A)\mu_B}{2(1+s)(\mu_A - \mu_B)}\left(\frac{\omega\varphi}{2}\right).\end{aligned} \tag{4..10}$$

For ISP_B $\Gamma_B(\mu_A, \mu_B)$ is defined as:

$$\begin{aligned}\Gamma_B(\mu_A, \mu_B) \equiv & \frac{1}{18(1+s)}\left(\varphi - \frac{\omega}{2}\right)^2 + \left(\frac{1}{3} - \frac{\mu_A - \mu_B}{9(1+s)}\right)\left(\varphi - \frac{\omega}{2}\right) + \\ & - \frac{(1-\lambda\mu_B)\mu_A}{2(1+s)(\mu_A - \mu_B)}\left(\frac{\omega\varphi}{2}\right).\end{aligned} \tag{4..11}$$

Equation (4..9) allows to compare the network neutrality and network discrimination equilibrium profits. In particular, ISP_A and ISP_B improve their profits by Γ_A and Γ_B, respectively. Both improvements depend on the expressions φ and $\frac{\omega}{2}$. As observed in equation (4..8), φ is the extra-discount that ISP_B gives to its users, by transferring a proportion of the fee paid by the CP buying the priority. In equation (4..6), we defined ω as the extra-disutility faced by users in the small network due to the higher congestion. Therefore, $\varphi - \frac{\omega}{2}$ represents the overall attractiveness gained by ISP_B because of the relative change of its end-user price with respect to network A. The improvement of ISP profits, given by in equations (4..10) and (4..11), can be decomposed in three parts:

- $\frac{1}{18(1+s)}\left(\varphi - \frac{\omega}{2}\right)^2$ captures the average improvement in total profits;

- the second part captures the extentto which profits change due to users:

  - In ISP_A, the term $-\left(\frac{1}{3}+\frac{\mu_A-\mu_B}{9(1+s)}\right)\left(\varphi-\frac{\omega}{2}\right)<0$. ISP_A is relatively less attrative with respect to ISP_B and, therefore, profits derived from users are less relevant.

  - In ISP_B, $\left(\frac{1}{3}-\frac{\mu_A-\mu_B}{9(1+s)}\right)\left(\varphi-\frac{\omega}{2}\right)>0$. ISP_B gains market share and users account for a larger share of profits;

- the last part captures how profits change due to the fee charged to the CP buying the priority.

  - In ISP_A, $\frac{(1-\lambda\mu_A)\mu_B}{2(1+s)(\mu_A-\mu_B)}\left(\frac{\omega\varphi}{2}\right)>0$. ISP_A transfers to users (in form of a discount) a lower fraction of the fee charged to the prioritized CP.

  - In ISP_B, $-\frac{(1-\lambda\mu_B)\mu_A}{2(1+s)(\mu_A-\mu_B)}\left(\frac{\omega\varphi}{2}\right)<0$. ISP_B applies a higher discount to attract users by transferring a larger part of the fee. Therefore, the fee charged to the prioritized CP has a lower direct contribution to the extra-profits of ISP_B.

Further details on profits derivation are provided in Appendix 4.8.

Long run analysis We investigate the case whether in the long run (i.e., when network capacity is endogenous) Internet Service Providers have higher incentive to invest in capacity expansion in a discriminating regime rather than in a neutrality regime. When discrimination is allowed, equilibrium profits are given by equation (4..9). In particular, when both networks prioritize one content, ISP_A's profits increase by Γ_A and ISP_B's

ones by Γ_B. Therefore, if Γ_j is increasing in μ_j (that is, if profits in network discrimination increases on one ISP capacity), then ISP_j has higher incentives to invest if it is allowed to discriminate.

Note that profits increment Γ_j depends on the change in relative ISPs' attractiveness $\varphi - \frac{\omega}{2}$. Remember that $\omega(\mu_A, \mu_B)$ is the function of capacities that denotes the difference in congestion disutilities between A and B, and that $\varphi(\mu_A, \mu_B)$ is the function of capacities that denotes the difference in the allowed user-price discount.

Lemma 4..1 *Functions $\omega(\mu_A, \mu_B)$ and $\varphi(\mu_A, \mu_B)$ are increasing in the capacity of the larger network (μ_A) and decreasing in the capacity of the smaller network (μ_B).*

This lemma states that when ISP_A expands its capacity, the disutility gap of its users and of ISP_B's users increases. Additionally, ISP_B needs to transfer a higher share to its users of the content fee to compensate for the utility loss. On the contrary, if ISP_B expands its capacity, the disutility gap diminishes and, therefore, the so does the need to compensate users. This condition is described in the following lemma.

Lemma 4..2 *$\varphi - \omega$ is an increasing function of the capacity of the larger network (μ_A) and a decreasing function of the capacity of the smaller network (μ_B).*

This result states that when ISP_A (ISP_B) expands its capacity, it increases its attractiveness relative to ISP_B (ISP_A).

Therefore, the overall effect of capacities expansion depends on the relationship between the extra-disutility created when discriminating and the margin that ISPs can extract from CP by charging a priority fee.

Proposition 4..8 *When ISPs are allowed to discriminate between content providers, and CPs gain a sufficiently high margin, both networks have less incentives to expand their capacities than in a neutrality regime.*

The intuition behind is that, on one hand, an ISP that invests in capacity increases its attractiveness since it provides a lower congestion to users, so it is able to increase its market share. On the other hand, higher capacity reduces congestion disutility to users, so less users switch to the priority CP.

Hence, if an ISP expands its capacity more than its competitor, it increases its total market share by attracting users of the other ISP that browse the discriminated content (since they are more penalized in that network). Overall, the ISP that invests more in capacity increases its share of users that browse the discriminated content, but at the same time it can only negotiate a lower fee with the CP buying the priority service. It is to note that If CP's margin is high enough compared to users' extra-disutility (which is reflected in lower market share or a lower subscription price), a network prefers to lose consumers and charge a high fee, since CP's willingness to pay is very high as well.

4.7 Policy implications and conclusion

The fact that there are cases when, under discrimination, both ISPs have less incentives to expand their network capacity compared to the neutrality regime, has important policy implications. Therefore, a regulator concerned with improving the quality of networks through the capacity should impose a neutrality regime. However, some considerations can be made concerning: the implications of the assumptions made;

and the policy goal of a regulator.

For our results, two assumptions are particularly relevant: the profitability of content providers and the initial asymmetry of ISP network capacity. We assumed that that both ISPs always find it profitable to discriminate. Our goal was to capture and analyze the most interesting cases. If this assumption is relaxed, that is, if CPs have low profit margins, then the charged fee must also be low enough and ISPs' revenues from users have greater importance when compared to revenues attained from CPs. Hence, ISPs may not choose to discriminate since the the loss of revenues from users who migrate to the other network (due to extra-disutility) is not compensated by the revenues earned from CPs. Hence, the outcome is the one of network neutrality.

If networks are symmetric, that is, ISPs have the same initial network capacity, then the results would change. ISPs discount an amount of the charged fee to users. The discount is allowed to all the users while the fee is only paid by the users who switch from the discriminated content to the prioritized content.[19] The overall effect in ISP profits is neutral, in the sense that profits do not change with respect to net neutrality. Therefore, the incentives to invest would be the same as in the neutrality regime. However, this result would only hold under the hypothesis of no binding consumers participation.

We have seen that, when asymmetric ISPs discriminate between contents, incentives to invest in network expansion are lower. This conclusion suggests that regulators need to impose network neutrality to achieve larger network capacities. However, the European position is not completely clear to this regard. The European Commission

[19] In more detail, in our model ISPs discount $\frac{2}{3}$ of the fee to all their users.

and the Body of European Regulators for Electronic Communications (BEREC) recognize the importance of developing faster, next generation networks. They claim that providing conditions for a competitive ISPs market allows achieving proper incentives to invest and, as a consequence, network neutrality.

In our model, a competitive market for ISPs is intended as a market where migrating from a ISP to another is costless to consumers since the transport cost when connecting to an ISP is always zero. However, in practice, horizontal differentiation between ISPs is impossible to wipe out. For instance, ISPs continuously introduce a package extra-services to theonline service (such as modems, TV programs, or phone calls). Hence, consumers face a cost if they cannot connect to the preferred Internet Service offer. Since each ISP has a captive market, the only way a CP can cover the whole market is by contract access through all the ISPs. This gives some degree of bargaining power to ISPs, which can extract part of the CP surplus by charging a fee on the provision of priority service.

A policy that limits ISPs' differentiation capacity and makes online service provision a competitive market could probably achieve network neutrality as market outcome. If consumers can freely migrate from an ISP to another, they would be rather unwilling to accept congestion disutility associated to discrimination. Hence, more discriminated consumers would migrate to the competitor when an ISP prioritizes a content. This would make discrimination very costly to ISP, meaning a great loss in terms of consumers. To this extent, the European perspective would be correct: warranting ISP competition could allow to achieve network neutrality without imposing it by law. However, if an ISP can recover the lower revenue due to consumers loss through

a high prioritization fee, network discrimination would still be profitable (and desirable from the ISP's point of view). In this case, as we have seen, incentives to invest in capacity expansion are lower.

Overall, the main concern of the European Commission should not only be to guarantee a competitive ISP market for users but also to provide a competitive ISP market for CPs. If ISPs have lower bargaining power when negotiating the priority fee, their capacity to extract CPs surplus would be lower as well. Therefore, the revenue from the prioritization fee would not compensate for the users loss due to discrimination. In this case, an ISP would prefer to not discriminate and compete for users by investing in larger network capacity.

A final remark can be done concerning the potential path for future research. While in our model the attention has been focused on the actions of ISPs, little was analyzed about the CP side of the market. We have seen that a low margin extraction of a CP associated with higher ISP competition implies that network neutrality is a more profitable regime for ISPs and can to higher investments in capacity expansion. However, a similar outcome occurs if the CP margin from advertising revenue is very small. Even though ISPs have a great bargaining power, if the margin they can extract is small, network neutrality is more profitable. Therefore, a regulator should also investigate CPs' advertising margin and negotiation, assessing if high margins (when present) are justified within the nature of the market or derive some market power practice.

4.8 Appendix

1. Maximization problem of the ISP in the neutral network:

$$\max_{p} \ \pi_M = p \ s.t. \ v - w - tx^* - p \geq 0.$$

2. Profit of the ISP in a discriminatory network:

$$\max_{\tilde{p}} \ \tilde{\pi}_M = \tilde{p} + f \qquad \text{s.t.} \quad v - w_1(\tilde{x}) - t\tilde{x} - \tilde{p} \geq 0.$$

3. Calculation of the fee by Nash bargaining:

$$\max_{f} \ [f - m_2(2\tilde{x} - 1)\lambda]^{\theta}[(m_1\tilde{x}\lambda - f) - m_1(1 - \tilde{x})\lambda]^{1-\theta} \Rightarrow f = [m_2 + \theta(m_1 - m_2)](2\tilde{x} - 1)\lambda$$

where $\theta \in [0, 1]$ denotes the ISP's bargaining power.

4. CP's profits under discrimination: when the ISP assigns the priority to CP_1, each content provider's profit will be respectively given by

$$\tilde{\pi}_1 = m_1\tilde{x}\lambda - [m_2 + \theta(m_1 - m_2)](2\tilde{x} - 1)\lambda; \qquad \tilde{\pi}_2 = m_2(1 - \tilde{x})\lambda.$$

5. First-order conditions of the problem of ISPs under the neutral network:

$$\frac{\partial \pi_A}{\partial p_A} = 0 \Longleftrightarrow \frac{1+s}{2} + \frac{\mu_A - \mu_B + p_A - p_B}{2(1+s)} - \frac{p_A}{2(1+s)} = 0$$

$$\frac{\partial \pi_B}{\partial p_B} = 0 \Longleftrightarrow \frac{1}{2} - \frac{\mu_A - \mu_B + p_B - p_A}{2(1+s)} - \frac{p_B}{2(1+s)} = 0$$

Proof of Proposition 4..1 Given the definitions of indifferent users, the difference between $\tilde{x}_B$ and $\tilde{x}_A$ is

$$\begin{aligned}\tilde{x}_B - \tilde{x}_A &= \frac{1}{2} + \alpha\frac{1-\lambda\mu_B}{2t\mu_B} - \left(\frac{1}{2} + \alpha\frac{1-\lambda\mu_A}{2t\mu_A}\right) \\ &= \alpha\frac{\mu_A(1-\lambda\mu_B) - \mu_B(1-\lambda\mu_A)}{2t\mu_A\mu_B} \\ &= \alpha\frac{\mu_A - \mu_B}{2t\mu_A\mu_B} > 0.\end{aligned}$$

Assumption 4.2 ($1 - \lambda\mu_A > 0$) guarantees that $\tilde{x}_A > x^*$.

Proof of Proposition 4..2

Proposition 4..2.a When both ISPs prioritize the same content, the difference between the disutility gaps in the two networks is $(w_{1,B}(\tilde{x}_B) + w_{2,B}(\tilde{x}_B)) - (w_{1,A}(\tilde{x}_A) + w_{2,A}(\tilde{x}_A))$, that is

$$\frac{\alpha^2(\mu_A - \mu_B)(\mu_B + \mu_A(1-2\lambda\mu_B))}{4(1+s)t\mu_A^2\mu_B^2} > 0.$$

Proposition 4..2.b When the ISPs prioritize different content providers, the difference between the disutility gaps in the two networks is $(w_{1,B}(\tilde{x}_B) + w_{2,B}(\tilde{x}_B)) - (w_{2,A}(1-\tilde{x}_A) + w_{1,A}(1-\tilde{x}$ that is

$$\frac{\alpha^2(\mu_A - \mu_B)(\mu_B + \mu_A(1-2\lambda\mu_B))}{4(1+s)t\mu_A^2\mu_B^2} > 0.$$

Proof of Proposition 4..3 The result of the comparison between indifferent users depends on the sign of the following difference:

$$\frac{\alpha(\mu_A + \mu_B - 2\lambda\mu_A\mu_B)}{4(1+s)\mu_A\mu_B} - \frac{\alpha(\mu_A - \mu_B)}{4(1+s)\mu_A\mu_B}$$

that is, on the sign of $\frac{\alpha(1-\lambda\mu_A)}{2(1+s)\mu_A}$, which is always positive (given our assumption $1 > \lambda\mu_A$).

Proof of Proposition 4..4 Let define $\tilde{\tau} \equiv \frac{\alpha(\mu_A - \mu_B)}{4(1+s)\mu_A\mu_B}$ and $\hat{\tau} \equiv \frac{\alpha(\mu_A + \mu_B - 2\lambda\mu_A\mu_B)}{4(1+s)\mu_A\mu_B}$. In the case both ISPs prioritize CP_1, total market share of ISP_A is

$$\tilde{\sigma}_A = \tilde{x}_A(y^* + \omega - \tilde{\tau}) + (1 - \tilde{x}_B)(y^* + \omega + \tilde{\tau}) + (y^* + \omega)(\tilde{x}_B - \tilde{x}_A)$$

$$= y^* + \omega + \tilde{\tau}(1 - (\tilde{x}_A + \tilde{x}_B)) = y^* + \frac{\omega}{4(1+s)}$$

given that $\tilde{\tau}(1-(\tilde{x}_A+\tilde{x}_B)) = -\frac{\omega}{2}$. In the case ISP_A prioritizes CP_2 and ISP_B prioritizes CP_1, total market share of A is

$$\hat{\sigma}_A = (1 - \tilde{x}_A)(y^* + \omega - \hat{\tau}) + (1 - \tilde{x}_B)(y^* + \omega + \hat{\tau}) + (y^* + \omega)(\tilde{x}_B - (1 - \tilde{x}_A))$$

$$= y^* + \omega + \hat{\tau}(\tilde{x}_A - \tilde{x}_B) = y^* + \frac{\omega}{4(1+s)}$$

given that $\hat{\tau}(\tilde{x}_A - \tilde{x}_B) = -\frac{\omega}{2}$.

Users gained by prioritized CP under ISP_B When ISP_B sells the priority service, the possible outcomes are represented in Figure 4..7. In particular, in Figure 4..7 (a) , the market share gained by CP_1 when it purchases the priority is represented by the hatched area. Horizontal lines denote the users of ISP_B who switch from CP_2 to CP_1. Vertical lines represent users who were browsing CP_2 under B and now browse CP_1 under A. The shaded area define the users who switch from A to B but not constitute a market gain for CP_1.

In Figure 4..7 (b), we represent the case where the priority is sold to CP_2. The meaning of the areas is the same as in the previous case.

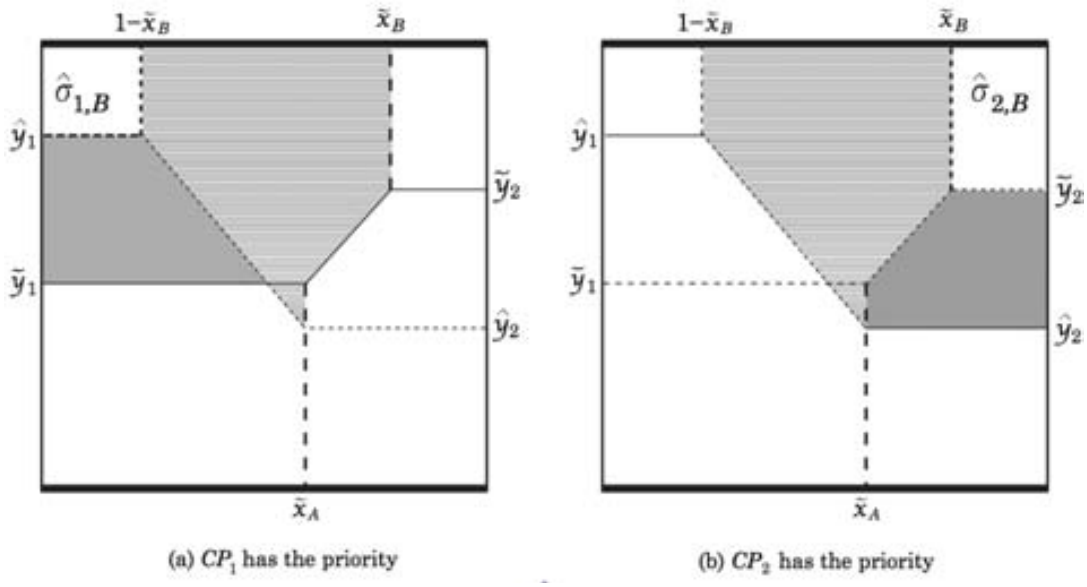

Figure 4..7: CP's user gain when prioritized by ISP_B

Proof of Proposition 4..5 When both ISPs sell the priority to CP_1, total market share of the content is

$$\tilde{\sigma}_1 = \tilde{x}_A + (2 - \tilde{y}_1 - \tilde{y}_2)\frac{\tilde{x}_A - \tilde{x}_B}{2} = \tilde{x}_B - \frac{\tilde{x}_B - \tilde{x}_A}{2}\left(y^* + \frac{\omega}{2(1+s)}\right).$$

If ISP_A sells the priority to CP_2, total market share of CP_1 would be $\tilde{x}_B + (1 - \tilde{x}_A - \tilde{x}_B)\left(y^2 + \frac{\omega}{2}\right)$. If ISP_B sells the priority to CP_2, total market share of CP_1 would be $1 - \tilde{x}_B - (1 - \tilde{x}_A - \tilde{x}_B)\left(y^2 + \frac{\omega}{2}\right)$. Total market share of CP_1 diminishes by $(2\tilde{x}_A - 1)\left(y^* + \frac{\omega}{2(1+s)}\right)$ if A deviates prioritizing 2. If B deviates, market share of 1 diminishes by $(2\tilde{x}_B - 1)\left(1 - y^* - \frac{\omega}{2(1+s)}\right)$. Recalling that $(2\tilde{x}_A - 1) = \frac{\alpha(1-\lambda\mu_A)}{t\mu_A}$ and $(2\tilde{x}_B - 1) = \frac{\alpha(1-\lambda\mu_B)}{t\mu_B}$, we get the market share variations in the proposition.

Proof of Proposition 4..6 The first order conditions/reaction functions for a maximum are

$$\frac{\partial \pi_A}{\partial p_A} = 0 \Longrightarrow$$
$$\frac{1}{2} + \frac{p_B - 2p_A + \mu_A - \mu_B}{2(1+s)} + \frac{\omega(\mu_A, \mu_B)}{2} - \frac{\alpha(1-\lambda\mu_A)(m_2 - \theta(m_2 - m_1))}{2(1+s)t\mu_A} = 0$$

$$\frac{\partial \pi_B}{\partial p_B} = 0 \Longrightarrow$$
$$\frac{1}{2} - \frac{2p_B - p_A + \mu_A - \mu_B}{2(1+s)} - \frac{\omega(\mu_A, \mu_B)}{2} - \frac{\alpha(1-\lambda\mu_B)(m_2 - \theta(m_2 - m_1))}{2(1+s)t\mu_B} = 0$$

The second order conditions are readily satisfied, $\frac{\partial^2 \pi_A}{\partial p_A^2} = \frac{\partial^2 \pi_B}{\partial p_B^2} = -\frac{1}{1+s} < 0$.

Proof of Proposition 4..7 Recall that

$$y^*(\tilde{p}_A, \tilde{p}_B) = \frac{1}{2} + \frac{p_B^* - p_A^*}{2(1+s)} + \frac{\mu_A - \mu_B}{2(1+s)} - \frac{1}{6(1+s)}\left(\omega + \frac{\alpha(\mu_A - \mu_B)}{t\mu_A\mu_B}\lambda(\theta m_1 + (1-\theta)m_2)\right)$$

$$y^*(\tilde{p}_A, \tilde{p}_B) = \sigma_A^* - \frac{1}{6(1+s)}\left(\omega + \frac{\alpha(\mu_A - \mu_B)}{t\mu_A\mu_B}\lambda(\theta m_1 + (1-\theta)m_2)\right).$$

Given our definition of φ and market shares defined in Proposition 4..4, we derive equilibrium market shares. The sign of the change in market shares with respect to the case of network neutrality is given by Assumption 4.5.

Equilibrium profits Note that equilibrium prices can be written as

$$\tilde{p}_A = p_A^* - \frac{1}{3}\left(\varphi - \frac{\omega}{2}\right) - \frac{\alpha(1-\lambda\mu_A)}{t\mu_A}\lambda(\theta m_1 + (1-\theta)m_2)$$
$$\tilde{p}_B = p_B^* + \frac{1}{3}\left(\varphi - \frac{\omega}{2}\right) - \frac{\alpha(1-\lambda\mu_B)}{t\mu_B}\lambda(\theta m_1 + (1-\theta)m_2).$$

Using the fact that $\sigma_A^* = \frac{p_A^*}{2(1+s)}$ and $\sigma_B^* = \frac{p_B^*}{2(1+s)}$, equilibrium market shares can be written as

$$\tilde{\sigma}_A = \sigma_A^* - \frac{\sigma_A^*}{3p_A^*}\left(\varphi - \frac{\omega}{2}\right)$$

$$\tilde{\sigma}_B = \sigma_B^* + \frac{\sigma_B^*}{3p_B^*}\left(\varphi - \frac{\omega}{2}\right).$$

Therefore, total revenues from users, $\tilde{p}_A\tilde{\sigma}_A$ and $\tilde{p}_B\tilde{\sigma}_B$, are

$$\begin{aligned}\tilde{p}_A\tilde{\sigma}_A =& p_A^*\sigma_A^* + \frac{1}{18(1+s)}\left(\varphi - \frac{\omega}{2}\right)^2 - \frac{2\sigma_A^*}{3}\left(\varphi - \frac{\omega}{2}\right) + \\ &- \frac{\alpha(1-\lambda\mu_A)}{t\mu_A}\left(\sigma_A^* - \frac{\sigma_A^*}{3p_A^*}\left(\varphi - \frac{\omega}{2}\right)\right)\lambda(\theta m_1 + (1-\theta)m_2)\end{aligned}$$

$$\begin{aligned}\tilde{p}_B\tilde{\sigma}_B =& p_B^*\sigma_B^* + \frac{1}{18(1+s)}\left(\varphi - \frac{\omega}{2}\right)^2 + \frac{2\sigma_B^*}{3}\left(\varphi - \frac{\omega}{2}\right) + \\ &- \frac{\alpha(1-\lambda\mu_B)}{t\mu_B}\left(\sigma_B^* + \frac{\sigma_B^*}{3p_B^*}\left(\varphi - \frac{\omega}{2}\right)\right)\lambda(\theta m_1 + (1-\theta)m_2)\end{aligned}$$

Proof of Lemma 8 The First-order derivatives are:

$$\frac{\partial\omega}{\partial\mu_A} = \frac{\alpha^2(1-\lambda\mu_A)}{t\mu_A^3} > 0 \qquad \frac{\partial\omega}{\partial\mu_B} = -\frac{\alpha^2(1-\lambda\mu_B)}{t\mu_B^3} < 0$$

$$\frac{\partial\varphi}{\partial\mu_A} = \frac{\alpha}{t\mu_A^2}\lambda(\theta m_1 + (1-\theta)m_2) > 0 \qquad \frac{\partial\varphi}{\partial\mu_B} = -\frac{\alpha}{t\mu_B^2}\lambda(\theta m_1 + (1-\theta)m_2) < 0.$$

Proof of Lemma 9 In particular, the derivatives of the difference $\varphi - \frac{\omega}{2}$ are

$$\frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial\mu_A} = \frac{\alpha}{t\mu_A^3}\left(\lambda(\theta m_1 + (1-\theta)m_2) - \frac{\alpha(1-\lambda\mu_A)}{2\mu_A}\right) > 0$$

$$\frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial\mu_B} = -\frac{\alpha}{t\mu_B^3}\left(\lambda(\theta m_1 + (1-\theta)m_2) - \frac{\alpha(1-\lambda\mu_B)}{2\mu_B}\right) < 0.$$

Proof of Proposition 4..8 The derivative of Γ_A with respect to μ_A is

$$\begin{aligned}\frac{\partial \Gamma_A}{\partial \mu_A} =& \frac{1}{9(1+s)}\left(\varphi - \frac{\omega}{2}\right)\frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial \mu_A} - \frac{\sigma_A^*}{3}\frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial \mu_A} + \\ & - \frac{1}{9(1+s)}\left(\varphi - \frac{\omega}{2}\right) + \frac{(1-\lambda\mu_A)\mu_B}{2(1+s)(\mu_A - \mu_B)}\frac{\partial\left(\frac{\omega\varphi}{2}\right)}{\partial \mu_A} + \\ & \frac{(1-\lambda\mu_B)\mu_B}{2(1+s)(\mu_A - \mu_B)^2}\left(\frac{\omega\varphi}{2}\right).\end{aligned}$$

Rearranging terms we get

$$\begin{aligned}\frac{\partial \Gamma_A}{\partial \mu_A} =& -\frac{1}{9(1+s)}\left(\varphi - \frac{\omega}{2}\right)\left(1 - \frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial \mu_A}\right) - \frac{\sigma_A^*}{3}\frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial \mu_A} + \\ & + \frac{\mu_B}{2(1+s)(\mu_A - \mu_B)}\left((1-\lambda\mu_A)\frac{\partial\left(\frac{\omega\varphi}{2}\right)}{\partial \mu_A} + \frac{(1-\lambda\mu_B)}{\mu_A - \mu_B}\left(\frac{\omega\varphi}{2}\right)\right).\end{aligned}$$

Given Assumption 4.5, we have

$$\frac{1}{9(1+s)}\left(\varphi - \frac{\omega}{2}\right) > \frac{\mu_B}{2(1+s)(\mu_A - \mu_B)},$$

while Lemmas 4..1 and 4..2 provide

$$\left(1 - \frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial \mu_A}\right) > \left((1-\lambda\mu_A)\frac{\partial\left(\varphi \cdot \frac{\omega}{2}\right)}{\partial \mu_A} + \frac{(1-\lambda\mu_B)}{\mu_A - \mu_B}\left(\frac{\omega\varphi}{2}\right)\right)$$

therefore, $\frac{\partial \Gamma_A}{\partial \mu_A} < 0$.

The same result is obtained when computing the sign of the derivative of Γ_B with respect to μ_B, given that

$$\begin{aligned}\frac{\partial \Gamma_B}{\partial \mu_B} =& \frac{1}{9(1+s)}\left(\varphi - \frac{\omega}{2}\right)\frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial \mu_B} - \frac{\sigma_B^*}{3}\frac{\partial\left(\varphi - \frac{\omega}{2}\right)}{\partial \mu_B} + \\ & - \frac{1}{9(1+s)}\left(\varphi - \frac{\omega}{2}\right) + \frac{(1-\lambda\mu_B)\mu_A}{2(1+s)(\mu_A - \mu_B)}\frac{\partial\left(\frac{\omega\varphi}{2}\right)}{\partial \mu_B} + \\ & \frac{(1-\lambda\mu_A)\mu_A}{2(1+s)(\mu_A - \mu_B)^2}\left(\frac{\omega\varphi}{2}\right).\end{aligned}$$

REFERENCES CITED

1. Armstrong, M., 2006, "Competition in Two-Sided Markets", Rand Journal of Economics, 37: 668–691.

2. Baake, P. & Mitusch, K., 2007, "Competition with Congestible Networks", Journal of Economics, 91 (2): 151–176.

3. Cheng, H. K., Bandyopadhyay, S., and Guo, H., 2011, "The Debate on Net Neutrality: A Policy Perspective", Information Systems Research, 22: 60–82.

4. Choi, J. P. and Kim, B. C., 2010, "Net Neutrality and Investment Incentives", RAND Journal of Economics, 41(3): 446-471.

5. Ellison, G. 2006, "Bounded Rationality in Industrial Organization", Econometric Society Monographs, 42: 142.

6. Hermalin, B. E. and Katz, M. L., 2007, "The Economics of Product-Line Restrictions with an Application to the Network Neutrality Debate", Information Economics and Policy, 19 (2): 215–248.

7. Kramer, J. and Wiewiorra, L., 2009, "Network Neutrality and Congestion Sensitive Content Providers: Implications for Service Innovation, Broadband Investment and Regulation", MPRA Paper 16655, University Library of Munich, Germany.

8. Musacchio, J., Schwartz, G., and Walrand, J., 2009, "A Two-Sided Market Analysis of Provider Investment Incentives with an Application to the Net-Neutrality Issue", Review of Network Economics, 8 (1): 3.

9. Mobius, M., 2001, "Death through Success: The Rise and Fall of Local Service Competition at the Turn of the Century", Technical report, mimeo.

10. Rochet, J-C. and Tirole, J., 2006, "Two-Sided Markets: A Progress Report", Rand Journal of Economics, 37 (3): 645–667.

11. Schuett, F., 2010, "Network Neutrality: A Survey of the Economic Literature", Review of Network Economics, 9 (2): 1–14.

12. Sluijs, J. P., 2010, "Network Neutrality between False Positives and False Negatives: Introducing a European Approach to American Broadband Markets", Federal Communications Law Journal, 62: 77–117.