

HANDBOOK OF GAME THEORY AND INDUSTRIAL
ORGANIZATION, VOLUME II



Handbook of Game Theory and Industrial Organization, Volume II

Applications

Edited by

Luis C. Corchón

Department of Economics, Universidad Carlos III de Madrid, Spain

Marco A. Marini

Department of Social and Economic Sciences, Università di Roma La Sapienza, Italy



Edward Elgar
PUBLISHING

Cheltenham, UK • Northampton, MA, USA

© Luis C. Corchón and Marco A. Marini 2018

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical or photocopying, recording, or otherwise without the prior permission of the publisher.

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

A catalogue record for this book
is available from the British Library

Library of Congress Control Number: 2017952214

This book is available electronically in the **Elgaronline**
Economics subject collection
DOI 10.4337/9781788112789

ISBN 978 1 78811 277 2 (cased)
ISBN 978 1 78811 278 9 (eBook)
ISBN 978 1 78811 279 6 (2 volume set)

Contents

<i>List of contributors</i>	vii
<i>Foreword by Eric Maskin</i>	ix

1 Introduction <i>Luis C. Corchón and Marco A. Marini</i>	1
--	---

PART I COLLUSION AND MERGERS

2 Horizontal mergers in oligopoly <i>Ramon Faulí-Oller and Joel Sandonís</i>	7
3 Collusive agreements in vertically differentiated markets <i>Marco A. Marini</i>	34
4 Cartels and leniency: Taking stock of what we learnt <i>Giancarlo Spagnolo and Catarina Marvão</i>	57
5 Assessing coordinated effects in merger cases <i>Natalia Fabra and Massimo Motta</i>	91

PART II CONTESTS

6 Contest theory <i>Luis C. Corchón and Marco Serena</i>	125
7 Endogenous timing in contests <i>Magnus Hoffmann and Grégoire Rota-Graziosi</i>	147

PART III SPECIAL TOPICS

8 Firm pricing with consumer search <i>Simon P. Anderson and Régis Renault</i>	177
9 Market structure, liability, and product safety <i>Andrew F. Daughety and Jennifer F. Reinganum</i>	225
10 Strategic delegation in oligopoly <i>Michael Kopel and Mario Pezzino</i>	248
11 Platforms and network effects <i>Paul Belleflamme and Martin Peitz</i>	286
12 Auctions <i>Ángel Hernando-Veciana</i>	318
13 Differential oligopoly games in environmental and resource economics <i>Luca Lambertini</i>	338
14 Intellectual property <i>Miguel González-Maestre</i>	367

vi *Handbook of game theory and industrial organization: applications*

15	Healthcare and health insurance markets <i>Pau Olivella</i>	394
16	The microeconomics of corruption <i>Roberto Burguet, Juan-José Ganuza and José G. Montalvo</i>	420

PART IV EXPERIMENTAL AND EMPIRICAL EVIDENCE

17	Experimental industrial organization <i>Jordi Brandts and Jan Potters</i>	453
18	Empirical models of firms' R&D <i>Andrés Barge-Gil, Elena Huergo, Alberto López and Lourdes Moreno</i>	475

Index

Contributors

Simon P. Anderson, Department of Economics, University of Virginia, USA and CEPR, UK

Andrés Barge-Gil, Department of Economic Analysis II, Universidad Complutense de Madrid, Spain

Paul Belleflamme, Université catholique de Louvain, CORE and Louvain School of Management, Belgium

Jordi Brandts, Institute for Economic Analysis (CSIC) and Barcelona GSE, Spain

Roberto Burguet, Institute for Economic Analysis (CSIC) and Barcelona GSE, Spain

Luis C. Corchón, Department of Economics, Universidad Carlos III de Madrid, Spain

Andrew F. Daughety, Department of Economics and Law School, Vanderbilt University, USA

Natalia Fabra, Universidad Carlos III de Madrid, Spain

Ramon Faulí-Oller, Department of Economics, University of Alicante, Spain

Juan-José Ganuza, Department of Economics, University Pompeu Fabra and Barcelona GSE, Spain

Miguel González-Maestre, Departamento de Fundamentos del Análisis Económico, Universidad de Murcia, Spain

Ángel Hernando-Veciana, Department of Economics, Universidad Carlos III de Madrid, Spain

Magnus Hoffmann, Institute of Economics, University of St. Gallen, Switzerland

Elena Huergo, Departamento de Fundamentos del Análisis Económico I, Universidad Complutense de Madrid, Spain

Michael Kopel, Department of Organization and Economics of Institutions, University of Graz, Austria

Luca Lambertini, Department of Economics, University of Bologna, Italy

Alberto López, Departamento de Fundamentos del Análisis Económico I, Universidad Complutense de Madrid, Spain

Marco A. Marini, Department of Social and Economic Sciences, Università di Roma La Sapienza, Italy and CREI, Italy

Catarina Marvão, SITE, Stockholm School of Economics, Sweden and University College Dublin, Ireland

Eric Maskin, Harvard University, USA

José G. Montalvo, Department of Economics, University Pompeu Fabra and Barcelona GSE, Spain

Lourdes Moreno, Departamento de Fundamentos del Análisis Económico I, Universidad Complutense de Madrid, Spain

Massimo Motta, ICREA-Universitat Pompeu Fabra and Barcelona GSE, Spain

Pau Olivella, Department of Economics, Universitat Autònoma de Barcelona, Spain

Martin Peitz, Department of Economics and MaCCI, University of Mannheim, Germany

Mario Pezzino, University of Manchester, UK

Jan Potters, Department of Economics and CentER, Tilburg University, the Netherlands

Jennifer F. Reinganum, Department of Economics and Law School, Vanderbilt University, USA

Régis Renault, THEMA Research Centre, Université de Cergy-Pontoise, France

Grégoire Rota-Graziosi, CERDI-CNRS, Université d'Auvergne, and FERDI, France

Joel Sandónis, Department of Economics, University of Alicante, Spain

Marco Serena, Max Planck Institute for Tax Law and Public Finance, Munich, Germany

Giancarlo Spagnolo, SITE, Stockholm School of Economics, Sweden, University of Rome “Tor Vergata” and EIEF, Italy and CEPR, UK

Foreword

The publication of this *Handbook*, bringing together game theory and industrial organization, is an occasion worth celebrating. After all, industrial organization (IO) – the study of how firms in a given market behave – was game theory’s first systematic application to economics, and the success of that application had much to do with giving game-theoretic ideas the prominent place they now have in the economics profession more generally.

Before game theory remade industrial organization in the 1970s, most IO analyses focused on two extreme but simple kinds of markets: perfectly competitive and monopolistic. In a perfectly competitive market, there are many small sellers (all selling the same kind of good) and many small buyers (“small” here means that the quantities sold by a seller and bought by a buyer are tiny compared with the totals for the market). One might guess that large numbers of traders would make analysis complicated, but they actually simplify matters. If each seller is small relative to the market, its own behavior can’t affect other sellers appreciably. So when figuring out what it should do, it needn’t worry about how the others anticipate it might behave – their anticipations aren’t relevant. In other words, a seller doesn’t have to be strategic (symmetrically, neither does a buyer). And consequently an economic analyst has a relatively easy job predicting the seller’s behavior as the solution to a simple profit-maximization problem. Indeed, because the seller is selling the same sort of good as all its competitors, it will take the good’s market price as given (if it chooses a higher price, it will have no customers – since they can get a perfect substitute for less; and it will be overwhelmed by customers if it chooses a lower price). In other words, the seller has no market power.

In a monopolistic market, by contrast, there is just one seller (I shall continue to assume throughout that there are many small buyers). Thus, as with perfect competition, the seller doesn’t have to worry about what other sellers are thinking about it – this time because there are no other sellers. And so, again, the seller’s optimization exercise as well as the analyst’s prediction exercise are quite straightforward (although the seller now does have market power; its own behavior determines the market price).

However, the intermediate case, oligopoly – where there is more than one seller, but not so many that a single seller has no effect on competitors – is more difficult. Think of the American automobile industry as it used to be, consisting primarily of General Motors (GM), Ford, and Chrysler. When GM worked out which models to manufacture, how many units of each model to produce, and what prices to set, it had to take into account what it anticipated Ford and Chrysler would do, and their actions depended on their forecasts about GM. Clearly, grappling with these anticipatory interactions between firms is essential to understanding the automobile industry. Yet such interactions are potentially very complex. Specifically, when an oligopolistic firm A tries to predict what its rival, firm B, will do, it must anticipate what B anticipates A will do, and what B anticipates A anticipates B will do, and so on. That’s why Nash equilibrium (Nash, 1950) was such a breakthrough: it cuts through this potentially infinite sequence of mutual anticipations.

A situation like the automobile industry can be modeled as game (more precisely, a “non-cooperative” game) in which the firms are players, a rule for how a player behaves constitutes its strategy, and players’ strategy choices jointly determine their payoffs. Nash proposed that a good prediction for how players will behave in such a game is that they will choose Nash equilibrium strategies: a configuration of strategies from which no individual player gains by deviating. If each player chooses a strategy to maximize its payoff given its anticipation of others’ strategies, then a Nash equilibrium is simply a fixed point of these optimizations. That is, in equilibrium, players’ anticipations about other players are correct and thus the infinite sequence of anticipations is circumvented.

Nash equilibrium was, without doubt, the central foundation on which the game-theoretic literature in industrial organization (and, later, other fields in economics) was erected (and it was for this contribution that John Nash shared the 1994 Nobel Memorial Prize in Economics). Nevertheless, Nash’s work had at least two important economic precursors.

First, Cournot (1838) and Bertrand (1883) analyzed particular instances of duopoly (an oligopoly in which there are just two firms in the industry) in a game-theoretic way, even though game theory wasn’t to be developed formally until the twentieth century. Indeed, both Cournot and Bertrand used what amounted to Nash equilibrium to make their predictions of how firms will behave. Still, remarkable though they are, Cournot’s and Bertrand’s highly stylized analyses lacked Nash’s great generality. Thus, the fact that they had far less influence than Nash (1950) is quite understandable.

The other notable pre-Nash development was monopolistic competition, whose literature was initiated by Chamberlin (1933) and Robinson (1933). Like an oligopoly, a monopolistically competitive market is intermediate between monopoly and perfect competition. And like an oligopolist, a monopolistically competitive firm has market power (normally because the good it sells is not a perfect substitute for other sellers’ goods). However, the firm is presumed to be too small to affect its rivals’ behavior, and so the strategic interactions of oligopoly are absent.

It may seem surprising that Nash’s work, rather than von Neumann and Morgenstern’s foundational volume, *Theory of Games and Economic Behavior*, published six years before Nash (1950), had the primary impact on the industrial organization literature. I suspect that von Neumann and Morgenstern (1944) failed to make much of a dent in economics because it is largely devoted to cooperative game theory, which studies games where players can enter into binding coalitions and which normally presumes that the coalition of all players (the grand coalition) forms. This sort of theory is, unfortunately, unsuited to most real-life markets, where typically the grand coalition does not form. Indeed, even if it does arise (the OPEC cartel in the oil market may have been a reasonable approximation of a grand coalition), IO theorists want to understand why this happens and how the coalition sustains itself; they do not usually take the grand coalition for granted, contrary to the presumption of cooperative game theory.

Another surprise is that once Nash’s paper appeared, another 20 years passed for game-theoretic work in any economic field – let alone in industrial organization – to take off; there was remarkably little game theory in economics in the 1950s and 1960s (one important exception was Schelling’s, 1960, use of game-theoretic ideas to illuminate international relations). Here again, I can only speculate on the reasons, but I conjecture that two important extensions of Nash – Harsanyi’s (1968) treatment of games of

incomplete information (in particular, his concept of Bayesian equilibrium) and Selten's (1965) treatment of intertemporal games (in particular, his concept of subgame perfect equilibrium) – needed to be understood and digested by economists before they could make good use of game theory in their work.

In any event, the big game-theoretic applications to IO in the 1970s generally involved multiple periods and/or incomplete information; there was a flood of papers on topics such as tacit collusion by oligopolists, market entry by new firms, and limit pricing and predation by incumbent firms, all of which drew heavily on innovations by Harsanyi (1968) and Selten (1965) (who both shared the 1994 Nobel with Nash).

By the early 1980s, game theory had been such a success in industrial organization that it started being used in political economy, international economics, finance, and other areas of economic theory. And at the close of that decade, there was scarcely a self-respecting economics department that didn't offer game theory as an important component of its curriculum.

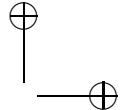
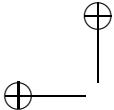
Industrial organization and game theory together led a revolution in economics. I am truly delighted that there is now a *Handbook* devoted to this transformative partnership.

REFERENCES

- Bertrand, J. (1883), "Review of Walras's Théorie Mathématique de la Richesse Sociale and Cournot's Recherches sur les Principes Mathématiques de la Théorie des Richesses", *Journal des Savants*, September, 499–508.
- Chamberlin, E. (1933), *The Theory of Monopolistic Competition: A Re-orientation of the Theory of Value*, Cambridge, MA: Harvard University Press.
- Cournot, A. (1838), *Recherches sur les Principes Mathématiques de la Théorie des Richesses*, Paris: L. Hachette.
- Harsanyi, J. (1968), "Games with Incomplete Information Played by Bayesian Players, I–III", *Management Science*, 14, 159–82, 320–34, 486–502.
- Nash, J. (1950), "Equilibrium Points in n-Person Games", *Proceedings of the National Academy of Sciences of the United States of America*, 36, 48–9.
- Robinson, J. (1933), *The Economics of Imperfect Competition*, London: Macmillan.
- Schelling, T. (1960), *The Strategy of Conflict*, Oxford: Oxford University Press.
- Selten, R. (1965), "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit", *Zeitschrift für die gesamte Staatswissenschaft*, 12, 201–324.
- von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Eric Maskin
Adams University Professor
Harvard University





1. Introduction

Luis C. Corchón and Marco A. Marini

In recent years, game theory has provided the ideal landscape in which to develop a wide range of industrial organization topics, as firms' entry, product differentiation, predation, delegation, mergers, collusion and R&D investment, auctions, health economics, contests and intellectual property rights, just to cite a few. Game theory was also recently and successfully applied to new fields such as law and economics, economics of networks, digital economy, experiments, corruption, and much much more.

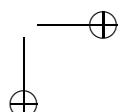
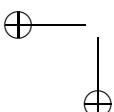
This second volume of the *Handbook* is devoted to presenting a wide set of applications of game-theoretic models to industrial organization topics. Thanks to the outstanding quality of the current contributors, this volume is relevant for both established researchers as well as to graduate and advanced undergraduate students.

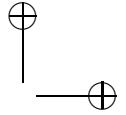
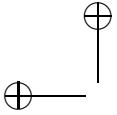
In this second volume we have organized the numerous applications into four distinct parts: (I) Collusion and Mergers; (II) Contests; (III) Special Topics; and (IV) Experimental and Empirical Evidence. More specifically, Part I of the *Handbook* deals with the analysis of horizontal mergers, with various collusive practices and mergers in vertically differentiated markets, with leniency effects in cartels and, finally, with coordinated effects in mergers. Part II introduces the state of the art in the literature dealing with both static and dynamic contests. Part III presents a quite wide spectrum of game-theoretic tools applied to various industrial organization topics such as firm pricing under consumer search, liability and product safety, strategic delegation, platforms and networks, auctions, intellectual property rights, healthcare markets, and corruption. All chapters use rigorous game-theoretic settings. Finally, Part IV presents two widely encompassing surveys of recent advances in experimental industrial organization and empirical models of firms' R&D. Let us now illustrate, in more detail, the content of all chapters composing this second volume of the *Handbook*.

PART I: COLLUSION AND MERGERS

In Chapter 2 on horizontal mergers in oligopoly, Ramon Faulí-Oller and Joel Sandonís analyze, in turn, the case of exogenous and endogenous mergers. In addition, the chapter examines in detail the models of horizontal mergers in vertically related industries and the models of mergers in an international setting, also looking at their welfare consequences.

In Chapter 3 on collusive agreements in vertically differentiated markets, Marco Marini introduces a number of game-theoretic tools that can be used to model the collusion between firms in a vertically differentiated market. The chapter starts reviewing the classical literature on two-firm collusion and, thus, the analysis is extended to a N-firm vertically differentiated market to study the incentive to form either a whole market cartel or partial cartels made up of subsets of firms colluding in prices. It is shown that a sufficient condition for the coalitional stability of the whole market cartel is the equidistance of firms' products along the quality spectrum. Also, adopting a standard infinitely repeated game approach, it is shown how an





2 *Handbook of game theory and industrial organization: applications*

increase in the number of firms in the market may have contradictory effects on the incentive of firms to collude. Finally, by means of a three-firm example, the feasibility of firm alliances under endogenous quality choice is studied.

In Chapter 4 on cartels and leniency, Giancarlo Spagnolo and Catarina Marvão study leniency policies. These policies reduce or cancel the sanctions for the first firm(s) that self-report being part of a cartel. They have become the main enforcement instrument used by competition authorities to fight against cartels. Hence, it is vital for competition authorities to understand how leniency programs affect firms' incentives, in order to optimize their design and administration. In this chapter, the authors review some of the key studies that have been undertaken to date, with emphasis on more recent contributions and highlight the main results and their limitations. The chapter concludes with a general assessment and an agenda for future research on this topic at the core of competition policy.

In the final chapter of Part I, Chapter 5 on assessing coordinated effects in merger cases, Natalia Fabra and Massimo Motta focus on the analysis of coordinated effects in merger cases; that is, the possibility that, after a merger, firms can increase their market power by coordinating their actions. The authors explain what coordinated effects are and how they can be assessed. For this purpose, they review the economic meaning of collusion, and assess the factors that allow firms to reach and enforce collusive outcomes. They also review some approaches for identifying and quantifying coordinated effects in practice, and provide an overview of the use of coordinated effects in European merger control.

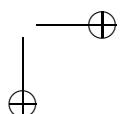
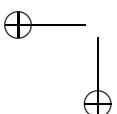
PART II: CONTESTS

In the first chapter of Part II, Chapter 6 on contest theory, Luis Corchón and Marco Serena introduce the readers to the main ingredients of contests, with a focus on how the efforts of the agents translate into probabilities of winning (i.e., the contest success function). In the second part of the survey, they focus on some extensions of the basic model, with a particular focus on dynamics, information and groups. They use the popular lottery model of contest with heterogeneous contestants to highlight their equilibrium properties and to review some results on how a designer can optimally design the contest.

In the second chapter of Part II, Chapter 7, Magnus Hoffmann and Grégoire Rota-Graziosi survey the extensive literature on endogenous timing in contests. They first introduce the structure of a two-player contest with either simultaneous or sequential moves and fixed prizes. They then present the case of ubiquitous contests with effort-dependent prizes, in which timing is endogenously determined. Finally, they conclude by looking in detail at the literature on sequential play, endogenous timing, and commitment in contests.

PART III: SPECIAL TOPICS

In the first chapter of this third part, Chapter 8 on firm pricing with consumer search, Simon P. Anderson and Régis Renault discuss in detail the basic concepts underpinning the theory of imperfectly competitive markets with consumer search. They first stress how the appropriate theoretical frameworks should involve sufficient heterogeneity among agents on both sides of the market. Moreover, they explain why the analysis of ordered search constitutes an



essential ingredient for modeling recent search environments. Finally, they examine in detail the important issue of the impact of the reduced search cost on prices, variety, product choice, and advertising practices.

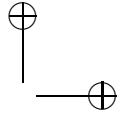
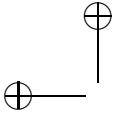
In Chapter 9, on market structure, liability, and product safety, Andrew Daughety and Jennifer Reinganum consider how models of imperfect competition provide insight into an important area of law: products liability, which is liability for harms and losses associated with goods and services sold via markets. Traditional law and economics analyses of products liability generally find no role for market structure or strategic interaction to influence safety or liability policy. Rather, different liability regimes, and alternative market structures, lead to the same private choice of safety, and this private choice is socially optimal. Daughety and Reinganum find that two simple (but plausible) model modifications, cumulative harm or endogenous fixed costs, yield a substantial impact of market structure on the choice of safety and liability regime.

In Chapter 10 on strategic delegation in oligopoly, Michael Kopel and Mario Pezzino provide the reader with a clear and intuitive description of the topic of strategic managerial incentives under oligopolistic competition. They describe in detail the related models of vertical separation, where a manufacturer delegates her decisions to a retailer, and the agent appointment game, where a principal delegates her decisions to a certain type of agent. Each of these themes is presented by discussing the seminal paper that first introduced the topic, its key assumptions and its applications along with some empirical and experimental evidence. The contributions that have provided important extensions to the basic frameworks are also discussed in the final section of the chapter.

In Chapter 11 on platforms and network effects, Paul Belleflamme and Martin Peitz review the key findings of the literature on network effects and two-sided platforms. They explore how to define network effects and markets with platforms, and investigate market demand and the provision of network goods. Then they outline the basic models of monopoly platforms and platform competition, and elaborate on some routes taken by recent research.

In Chapter 12 on auctions, Ángel Hernando-Veciana provides an overview of the advances in the auction field that have taken place in the last decade. To this aim, the survey starts with an introductory section in which the basic tools of analysis are summarized. Next, the main advances in three innovative areas of auction theory are spelled out in detail: position auctions, Internet auctions and combinatorial auctions. The final section of the chapter summarizes the major contributions to auction theory organized by topics.

In Chapter 13 on differential oligopoly games in environmental and resource economics, Luca Lambertini offers a comprehensive overview of the literature based on differential games whose main focus is the interplay between either regulated or unregulated oligopolistic firms' profit incentives and the preservation of the stock of natural capital. The first section introduces Cournot oligopoly games with either polluting emissions or resource extraction under open-loop rules, without regulation on emissions or access to the commons. The second section reviews the literature on environmental games with feedback structures, where firms may be subject to emission taxes and possibly activate R&D projects for green technologies. The third section considers games with exploitation of renewable and nonrenewable resources. The final sections are devoted, in turn, to corporate environmentalism and the Porter hypothesis as well as to the issue of international trade and the environment, both crucial for the ongoing debate on globalization and climate change.



4 *Handbook of game theory and industrial organization: applications*

In Chapter 14 on intellectual property, Miguel González-Maestre discusses the current literature on intellectual property rights from a perspective taking into account two main features of the evolution of modern economies: (1) the increasing level of complexity associated with the production and design of goods; and (2) the rapid development of new technologies of information and communication. To this end, the chapter mainly focuses on the recent literature dealing with the role of technological changes on the optimal design of patents and copyrights. This overview suggests that the substantial changes observed in Western economies, aiming at reinforcing intellectual property rights, cannot be justified either theoretically or empirically on the grounds of welfare or of creative and innovational incentives. Instead, alternative explanations based on rent seeking and lobbying activities by copyright or patent holders emerge as their most plausible rationale.

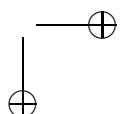
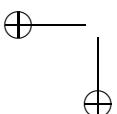
In Chapter 15 on healthcare and health insurance markets, Pau Olivella introduces some modeling tools for the analysis of healthcare provision and health insurance. In particular, the chapter devotes great attention to a series of topics for which the tools of industrial organization and game theory have proven most fruitful: (1) firms' incentives to invest in R&D in the pharmaceutical industry; (2) risk selection and screening of consumers; and (3) the effect of hospitals' competition.

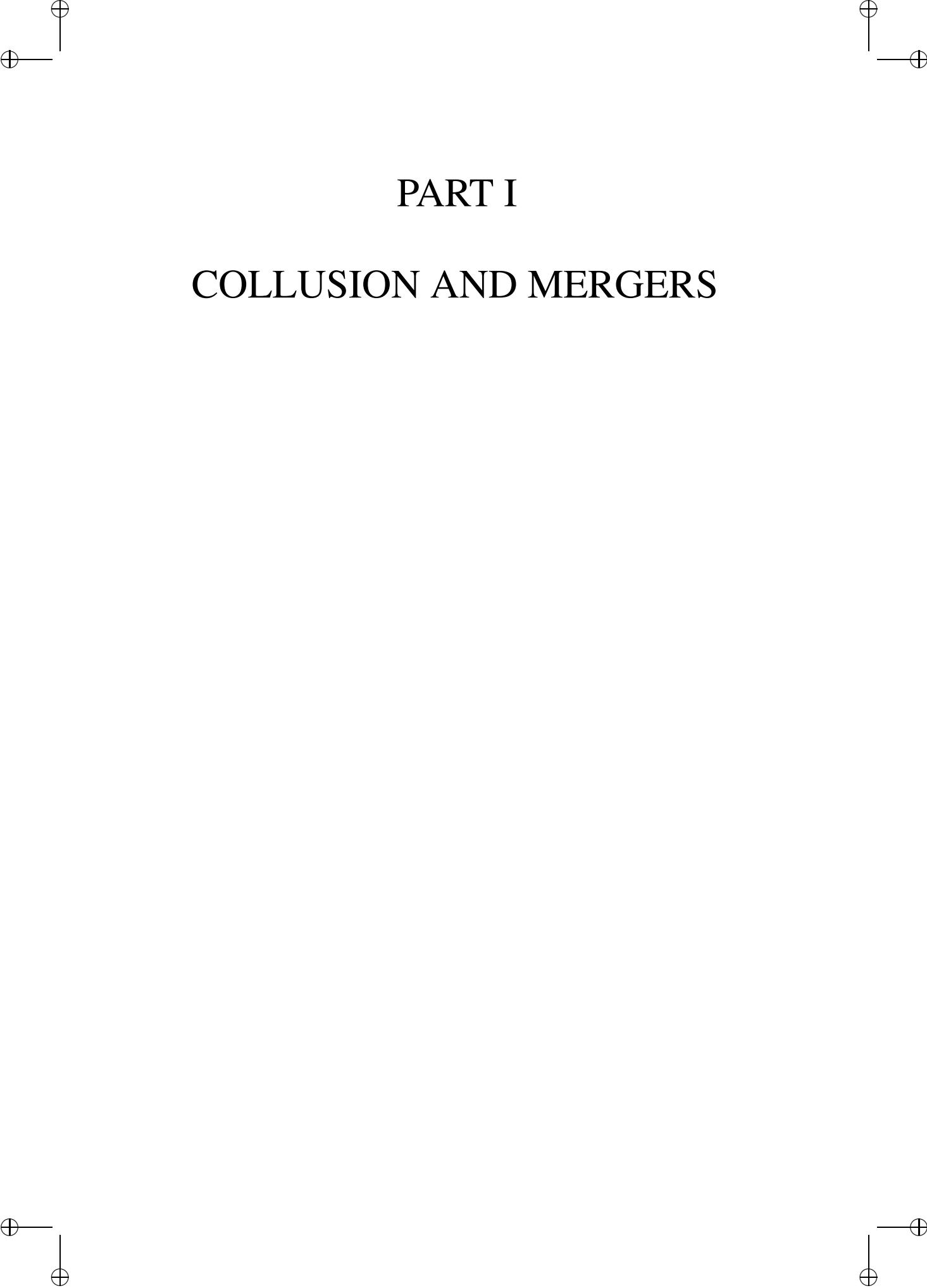
In Chapter 16 on the microeconomics of corruption, Roberto Burguet, Juan-José Ganuza and José G. Montalvo review the most recent research on corruption. They start by analyzing the seminal models of corruption built on three-tier delegation models. Then, they discuss the case of corrupted deals to see which main economic factors affect corruption. Incentives and compensations in bureaucracies, as well as the strict interplay of market and bureaucracies are discussed. Competition and contract design are also reviewed in relation to procurement in presence of corruptible agents. After reviewing theory, the authors turn to empirical evidence. Finally, they critically evaluate several anti-corruption mechanisms proposed by the literature to both control and eliminate illegal activities.

PART IV: EXPERIMENTAL AND EMPIRICAL EVIDENCE

In the first chapter of this fourth part of the *Handbook*, Chapter 17 on experimental industrial organization, Jordi Brandts and Jan Potters present a selective survey of the recent experimental studies on industrial organization issues. The first section of the chapter presents, starting with the classical models of Cournot, Bertrand and Stackelberg, the results of experiments based on static models involving the choices of quantities and prices. The second section deals with tacit collusion. The third section covers horizontal product differentiation and the fourth discusses experience and credence goods. The last section presents a few studies on entry deterrence and R&D competition.

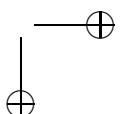
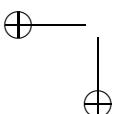
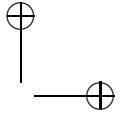
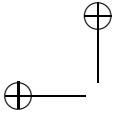
In the concluding chapter of the volume, Chapter 18, devoted to the empirical models of firms' R&D, Andrés Barge-Gil, Elena Huergo, Alberto López, and Lourdes Moreno survey the ever-growing empirical literature of R&D. They explain that this literature is still growing due to the increasing availability of micro-data. Taking this fact into account, the main purpose of this chapter is to provide an overview of three important topics covered by the recent literature: the determinants of firms' R&D investment, the link between R&D, innovation and productivity, and the analysis of the R&D black box. This chapter is presented as an invitation to industrial organization practitioners, both theorists and applied, to cross the bridge (and to change sides) between theory and applications.

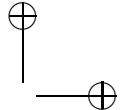
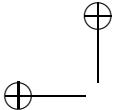




PART I

COLLUSION AND MERGERS





2. Horizontal mergers in oligopoly

*Ramon Faulí-Oller and Joel Sandonís**

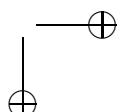
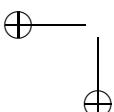
1 INTRODUCTION

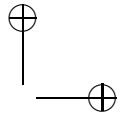
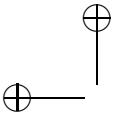
Mergers are a very important empirical phenomenon and form an important part of firms' strategies: we see daily merger announcements in the business press. There can be several different motives for mergers. Apart from an attempt to gain market shares and market power, mergers can seek efficiency gains through the combination of different assets, or by getting access to new technologies and know-how; they can also seek to reach new groups of consumers or new geographical markets, showing a desire for diversification; in other cases, mergers can also seek to safeguard access to important inputs. When studying mergers, we have to be cautious because there are different types of mergers that require different theoretical treatments. The first distinction is between conglomerate, horizontal and vertical mergers. In this chapter, we focus on horizontal mergers, that is, mergers between firms that are direct competitors. Conglomerate mergers are mergers that take place between firms operating in unrelated markets, whereas vertical mergers occur between firms operating at successive stages of the production process.

Horizontal mergers reduce competition. Mergers whose only aim is to reduce competition (or to gain market power) create the following free-riding problem. The reduction of competition due to a merger works as a public good that benefits all the firms in the industry, while the costs are only supported by the merging parties. This implies that non-merging firms benefit more from the merger than merging firms. Therefore, firms want competitors to merge. This result does not depend on the type of competition being either Bertrand or Cournot, however, the problem is more acute in a Cournot setting, where mergers may turn out to be unprofitable, i.e. the joint profits of the merging firms before the merger are higher than their joint profits after the merger. The explanation is the aggressive response of outsiders, which react to the output reduction by the merging firms by increasing their own output (quantities are strategic substitutes). This point was first raised by Salant, Switzer and Reynolds (1983). This response is so unexpected that it is usually known as the "merger paradox". It originated a huge literature that tried to obtain profitable mergers in a Cournot setting. For example, profitability of mergers increases if either demand or costs become more convex.

The free-rider problem also explains that profitable mergers do not materialize in the equilibrium of a non-cooperative merger game where mergers take place endogenously. For example, in Kamien and Zang (1990), in order to monopolize the industry a firm must pay to each competitor its outside option, which amounts to the duopoly profits, i.e. the profits that each of them would obtain if rejecting the offer, provided that the rest of the firms accept it. And this is so expensive that monopolization can only occur in equilibrium in very concentrated

* The authors acknowledge financial support from the Spanish Ministerio de Economía y Competitividad (ECO2015-65820-P) and from Generalitat Valenciana grant Prometeo/2013/037. The chapter was written while Faulí-Oller was visiting the Institut d'Anàlisi Econòmica (CSIC).





8 *Handbook of game theory and industrial organization: applications*

markets. On the other hand, the free-rider problem is attenuated in dynamic games. In this case, there is an additional incentive to merge today in order to induce new mergers in the future (Pesendorfer, 2005). In other dynamic settings, mergers are shown to be profitable as an alternative to either exit or internal investment (Gowrisankaran, 1999). More concentration is also obtained if cooperative merger games are considered. In this case, the merger process is not fully described and the focus is on the stability concept used to check whether different market structures can be the outcome of the merger game because no firm (or group of firms) have incentives to deviate.

Another strand of the literature considers horizontal mergers taking place in vertically related markets. In this setting, apart from reducing competition, mergers may also affect the terms of trade between upstream and downstream firms, which ultimately affect the efficiency of the latter firms and the result of the merger both in terms of profitability and social welfare.

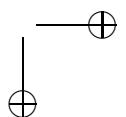
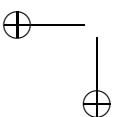
International mergers are also a very relevant empirical phenomenon. For example, the ratio of the value of global cross-border mergers to the value of global FDI in the three years before the global economic crises (2004–07) was very high, sometimes as high as 80 percent.¹ In the international mergers section, we study the profitability of cross-border mergers. It will depend on a parameter t that accounts for the difference between the cost of exporting and the cost of national production. Therefore, t includes both tariffs and transportation costs. It is shown that reductions in t may stimulate cross-border mergers.

Last, we study the social welfare consequences of horizontal mergers. This is an important issue because mergers affect all of us, regardless of whether we are consumers, entrepreneurs, regulators or policymakers. And it is especially important for competition authorities. Nowadays, most countries have laws and regulations that call for the antitrust authorities to engage in merger control. In the USA, for example, merger control has a long tradition, starting with the Clayton Act of 1914. According to this act, mergers that lead to a substantial lessening of competition are forbidden. In the European Union, however, merger control was introduced in 1990, and then revised in the 2004 Horizontal Merger Guidelines.

We will start with the result showing that mergers for market power, which increase price, may still increase social welfare (see Farrell and Shapiro, 1990 and McAfee and Williams, 1992). The reason is that although production is reduced as a consequence of the merger, it may be more efficient as long as a fraction of the output is shifted from inefficient firms to more efficient firms after the merger. Apart from increasing market power, a merger may generate efficiency gains. As far as social welfare is concerned, this merger generates a trade-off: on the one hand it reduces competition (which is bad) but, on the other hand, it reduces costs (which is good). This trade-off was first raised by Williamson (1968). The higher the efficiencies the more profitable mergers will be. If they are high enough, the merging firms gain more than non-merging firms from the merger. In fact, outsiders may lose as a result of the merger.

Assessing the welfare impact of a merger is very important but it ignores the fact that the antitrust authority has the opportunity to decide on a merger case only if it has been previously proposed by the merging firms. Therefore, it is more appropriate to model the interaction between firms and the antitrust authority as a game where both firms and the antitrust authority play strategically. The main feature of this game will be that merging firms

¹ In the six years since the start of the crises, however, this ratio has declined even below 60 percent (see the OECD report at <http://www.oecd.org/daf/inv/FDI-in-Figures-April-2014.pdf>).



have better information about merger characteristics than the antitrust authority. Merging firms may have better information about the cost savings induced by a given merger or about the set of possible mergers. In this setting, the government may prefer the antitrust authority to enforce a merger rule that is different from the one that maximizes the preferences of the government. In particular, if the government maximizes total welfare (the unweighted sum of consumer surplus and profits) it will ask the antitrust authority to follow a merger rule that gives more weight to consumer surplus than to profits. The reason is that this bias towards consumer surplus serves to counterbalance the fact that mergers are proposed by firms, which only care about profits.

Due to lack of space, there are several topics related to horizontal mergers that are not included in this chapter. First, we do not consider the effect that mergers may have on collusion (see, for example, Compte, Jenny and Rey, 2002 and Vasconcelos, 2004). In other words, we consider the unilateral effect of mergers while ignoring their possible coordinated effects. Second, we do not study how mergers may influence the decisions of firms on other strategic variables, like, for example, R&D. There is a huge literature on this topic, starting with the seminal paper by D'Aspremont and Jacquemin (1988).

The rest of the chapter is organized as follows. In the next section we start with the case of exogenous mergers and, in Section 3, we deal with endogenous mergers. In Section 4 we consider horizontal mergers in vertically related industries. Mergers in an international setting are studied in Section 5. Section 6 considers the social welfare effects of mergers. Finally, we conclude in Section 7.

2 EXOGENOUS MERGERS

In this section, we study the profitability of a given merger taken in isolation. A merger is said to be profitable if the profits of the merging firms increase relative to their combined profits in the status quo. This issue is interesting because, although at first sight it might seem surprising, there are cases where mergers are not profitable. In the endogenous merger section, we will analyze which mergers do occur in equilibrium in different merger games. We will see that profitability is a necessary but not sufficient condition for a merger to materialize. The reason is that participation in a merger is not attractive, because firms not participating in a merger obtain higher profits than merging firms. As Stigler (1950, pp. 25–26) put it “the promoter of a merger is likely to receive much encouragement from each firm – almost every encouragement, in fact, except participation”.

2.1 Mergers in Bertrand and Cournot Competition

The overall effect of a merger on profitability can be understood as the sum of two different effects: the first comes from the fact that the merging firms change their strategies and the second from the fact that the outsiders also change their strategies. The former effect can only be positive because firms are maximizing joint profits after the merger, while they were maximizing individual profits before the merger. So this effect increases the joint profits. The sign of the effect coming from the change in the strategy of outsiders is uncertain and depends crucially on whether strategies are strategic complements (Bertrand) or strategic substitutes (Cournot).

10 Handbook of game theory and industrial organization: applications

Under Bertrand competition, the merging firms increase their price after the merger to maximize joint profits. In order to illustrate this idea, suppose that firms i and j sell respectively substitutive goods i and j before the merger, with demands $D_i(p)$ and $D_j(p)$ and marginal production costs c_i and c_j . After the merger, their joint profits are given by:

$$\begin{aligned}\pi &= (p_i - c_i)D_i(p) + (p_j - c_j)D_j(p). \\ \frac{\partial\pi}{\partial p_i} &= D_i(p) + (p_i - c_i)\frac{\partial D_i(p)}{\partial p_i} + (p_j - c_j)\frac{\partial D_j(p)}{\partial p_i} = 0\end{aligned}$$

The third term in the first-order condition (FOC) would not appear if firm i had not merged with firm j and represents the idea that the merged firms internalize the externality that they impose on each other before the merger. This term is positive, because the goods are substitutes, and it pushes the optimal price upwards.

As we have strategic complements (best responses are upward sloping), the outsiders to the merger react by increasing their prices as well, which increases the insiders' demands and their joint profits. The resulting changes in the prices of both insiders and outsiders increase profits (Deneckere and Davidson, 1985, Braid, 1986 and Levy and Reitzes, 1992).

Under Cournot competition, however, the merging firms reduce their output to maximize joint profits. For example, suppose that firms i and j merge and $P_i(q)$ and $P_j(q)$ represent the inverse demand of each good. In this case, the joint profits of the merged firms are given by:

$$\begin{aligned}\pi &= q_i(P_i(q) - c_i) + q_j(P_j(q) - c_j). \\ \frac{\partial\pi}{\partial q_i} &= P_i(q) - c_i + q_i\frac{\partial P_i(q)}{\partial q_i} + q_j\frac{\partial P_j(q)}{\partial q_i} = 0.\end{aligned}$$

The third term in the FOC would not exist if firm i had not merged with firm j . It is negative and it pushes the optimal quantity downwards.

Then, as we have strategic substitutes, the outsiders will react by increasing their output. This reduces the profits of the merging firms because it reduces the price at which they sell their goods. Summarizing, the effect of a merger on profits under Cournot competition is uncertain because it is the sum of a positive and a negative effect. The negative effect will be small if most of the firms participate in the merger because in that case the number of outsiders will be small and their output reaction will also be small. Salant et al. (1983) show that with linear marginal costs, linear demand, homogeneous goods and symmetric firms, if the merging firms have a market share lower than 80 percent, the merger is not profitable. In Figure 2.1, we plot for every possible number of firms in the market, the minimal market share for a merger to be profitable. The function has a minimum of 80 percent in $n = 5$.

Given the surprising result obtained under Cournot competition, a great number of papers have tried to increase the profitability of mergers by changing the assumptions of the original model. As the problem hinges on the fact that outsiders react by increasing their outputs, the first group of papers has introduced new assumptions trying to reduce the reaction of non-merging firms. The second group of papers increases the profitability of the merging firms by introducing some efficiency gains. The third group explicitly considers the separation between ownership and management. Let's analyze each of these groups in more detail.

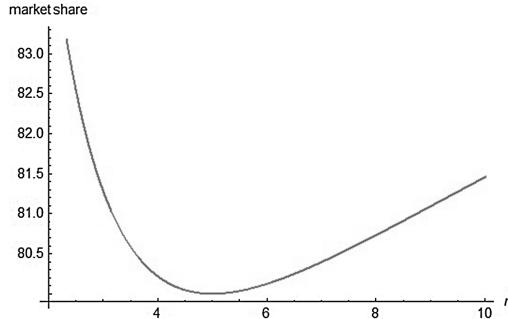


Figure 2.1 The minimal market share for a merger to be profitable for the case of linear demand

2.1.1 Reducing the reaction of outsiders

Convex costs Instead of linear costs, Perry and Porter (1985) and McAfee and Williams (1992) consider quadratic costs, $C(q) = dq^2$. Under this cost configuration, the greater is d the flatter the reaction curve and therefore, the lower the reaction of outsiders, which results in more profitable mergers. Observe that the slope of the reaction function is given by:

$$r'_i(q_{-i}) = -\frac{1}{2(1+d)}.$$

This result can be reinterpreted in terms of product differentiation because if profits are rewritten appropriately, d stands for the degree of product differentiation (Vives, 2002). The profits in Perry and Porter's (1985) model are written as:

$$(\alpha - q_i - q_{-i})q_i - dq_i^2.$$

Grouping terms we have:

$$(\alpha - (1+d)q_i - q_{-i})q_i.$$

But the last expression is simply the profit of a firm selling a differentiated product with no cost. Then, we have that the greater the product differentiation (the higher d) the higher the profitability of mergers. This result is a direct implication of Perry and Porter (1985) and it was formally stated by Lommerud and Sorgard (1997).

Convex demands If Perry and Porter (1985) changes the costs from the original model, Faulí-Oller (1997) changes the shape of the demand function to take the following form: $P(Q) = A - \frac{1}{b+1}Q^{b+1}$, where b parametrizes the degree of concavity of demand. The lower is parameter b the more convex is demand. And the lower is b the flatter is the reaction function

12 Handbook of game theory and industrial organization: applications

of firms and the more profitable mergers will be. Notice that the slope of the reaction function is given by:

$$r'_i(q_{-i}) = -1 / \left(\frac{1}{1 + bs_i} + 1 \right),$$

where s_i stands for firm i 's market share. If b increases, the slope in absolute value also increases.

We can check this result in Figure 2.2. It plots the minimal market share for a merger to be profitable for the case of unit elasticity demands. It can be easily seen that the minimal market share decreases and, therefore, there are some mergers that were unprofitable in the linear demand case and become profitable once we consider a more convex demand. With unit elasticity demands we have that mergers with a pre-merger market share lower than 50 percent are not profitable. This is the threshold identified by Cheung (1992).

Sequential mergers Mergers in Salant et al. (1983) are not profitable because outsiders react by increasing their output. This negative effect decreases as the number of non-merging firms reduces. Therefore a merger, by reducing this number, may induce new profitable mergers. Consider the following game with unit elasticity demands. Suppose we have four firms that play the following merger game. Firms 1 and 2 decide together whether they want to merge (M) or not (NM). Firms 3 and 4 do the same. Profits of firms as a function of the number of competitors n is given by $\pi(n) = \frac{1}{n^2}$. We have the following payoff matrix:

		Firms 3&4			
		NM	M	NM	M
Firms 1&2	$NM & M$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{9}$	$\frac{1}{9}$
	M	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{4}$	$\frac{1}{4}$

There are two Nash equilibria: either no firm merge (NM) or both firms merge (M). Then, firms want to merge if the rivals do (Faulí-Oller, 2000, Fumagalli and Vasconcelos, 2009 and Salvo, 2010). This result nicely illustrates the empirical evidence that mergers occur in waves (Andrade, Mitchell and Stafford, 2001).

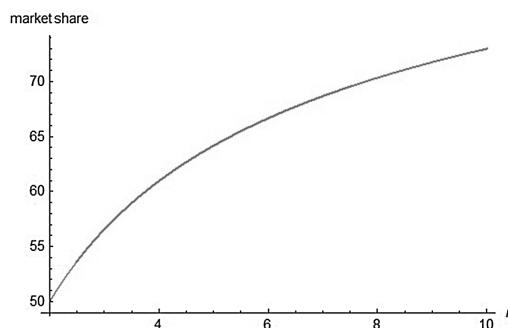


Figure 2.2 The minimal market share for a merger to be profitable for the case of unit elasticity demands

2.1.2 Introducing efficiency gains

So far, the only aim of the mergers that we have considered is to reduce competition. We have seen that, in Cournot settings, this effect is too weak to make mergers profitable. Next, we are going to consider cases where mergers allow firms to obtain efficiency gains, with the aim of stimulating merger profitability. Following Farrell and Shapiro (1990), it is convenient to distinguish between efficiencies obtained through synergies from efficiencies obtained without synergies. In the former case, mergers shift the cost function of the merging firms downwards. In the latter, mergers do not change the cost function of firms. Nevertheless, compared with the pre-merger situation, the merger can reduce costs by reallocating production from high-cost to low-cost merging partners. Motta and Vasconcelos (2005) study mergers with synergy gains by considering that the merger reduces the (constant) marginal cost of firms. They obtain that if the cost reduction is high enough mergers are not only profitable but they may also reduce price.

Falvey (1998) considers a model without synergy gains. He assumes that firms have constant marginal costs, but they may be asymmetric. When two firms with different costs merge, the merged entity produces at the cost of the most efficient firm. We can imagine that firms have excess capacity and they allocate all the production to the low-cost plant.

With linear demand, the merger of two firms (i is the efficient firm and j the inefficient) is profitable if:

$$\frac{s_i}{s_j} > \frac{n}{2} - \frac{1}{2n}, \quad (2.1)$$

where s_i denotes firm i 's market share and n the total number of firms. Using the FOC, it is possible to write the left-hand side of the previous inequality as:

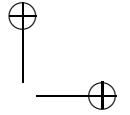
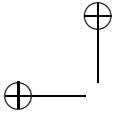
$$\frac{s_i}{s_j} = \frac{P - c_i}{P - c_j},$$

where P is the price and c_i firm i 's marginal cost. We can see that market share and costs are negatively correlated. Therefore, the merger will be profitable if the cost differential between the merging partners is high enough. In other words, if the efficiency effect is high enough. Observe that as the right-hand side of (2.1) is greater than 1 (except when $n = 2$), the merger of two symmetric firms is not profitable.

If demand decreases, price also decreases and the previous ratio increases. Then, it is more likely that the profitability condition is satisfied (observe that the right-hand side of (2.1) remains constant). So we can conclude that reductions in demand increase the profitability of mergers (Faulí-Oller, 2002). There is a long controversy on whether mergers occur in booms or in crises. The model seems to demonstrate that mergers occur in periods of declining demand as a way to rationalize production (Dutz, 1989 and Filson and Songsamphant, 2005).

2.1.3 Considering the separation between ownership and management

González-Maestre and López-Cuñat (2001) make the same assumptions as Salant et al. (1983) in terms of costs, demand and type of competition, but they assume that the owners delegate output decisions to managers, whose incentive package includes both profits and sales (Sklivas, 1987). The weight given to sales is chosen by owners and, in equilibrium,



14 *Handbook of game theory and industrial organization: applications*

it is positive in order to implement an aggressive market strategy. Therefore, price is lower with delegation than without delegation. They obtain that mergers are more profitable in this setting, namely, that, given the number of firms, the minimal market share for a merger to be profitable is lower with delegation than without delegation. Ziss (2001) obtains the same result by considering demands with a constant degree of concavity.

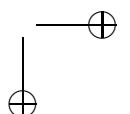
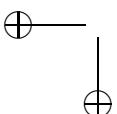
In this line, Faulí-Oller and Motta (1996) consider a model with three firms. Two are profit-maximizing firms and the third one is different because its owners delegate both market and merger decisions to a professional manager. The incentive scheme of the manager is similar to the one in González-Maestre and López-Cuñat (2001). In this case, there is an additional incentive for owners to increase the weight given to sales: the higher the weight the lower the price to be paid for a competitor. As in González-Maestre and López-Cuñat (2001) and Ziss (2001), they obtain that delegation stimulates the occurrence of mergers.

2.2 Other Strategic Variables

So far, we have studied situations where the only decision of firms referred either to price or to quantity. In this section we consider the case where apart from price or quantity firms decide on how many varieties of a good to produce. In particular, we consider Lommerud and Sorgard (1997), which deals with the case of multi-brand firms (for a more detailed explanation of this paper see Chapter 4 in Brito and Catalao-Lopes, 2006). They study a case where firms can choose the number of varieties of a good they offer and how this decision is affected by mergers. Varieties are modeled as differentiated goods with a linear and symmetric structure as in Deneckere and Davidson (1985). To produce a variety, firms incur a fixed cost F . There are three firms (1, 2, 3) and the authors study the profitability of a merger between firms 1 and 2. Pre-merger, given the cost per variety, it is optimal for each firm to produce only one variety. Post-merger, three different situations may arise: in Regime 1, the merged firm eliminates one variety; in Regime 2, there is no change in the varieties offered by firms; and finally, in Regime 3, firm 3 adds a new variety. Merger profitability in this setting depends on whether firms compete in quantities or in prices. With quantities, the merger in Regime 3 is never profitable: the negative effect on profitability of the increase in the production of non-participating firms is amplified in this case because firm 3 introduces a new variety. In Regime 2, the number of varieties does not change and the merger is profitable if varieties are differentiated enough, so that the increase in the output of firm 3 is not so important. In Regime 1, the merger is profitable if F is high enough and the degree of differentiation is high enough. With competition à la Bertrand, in Regime 3 the merger is never profitable. This is the most interesting case. In Deneckere and Davidson (1985), mergers are always profitable with price competition. In the present chapter, they are not profitable if the merger induces firm 3 to increase the number of varieties it offers. In Regime 2, we have the same situation as in Deneckere and Davidson (1985) so that the merger is profitable. In Regime 1, the merger is profitable because it reduces competition.

3 ENDOGENOUS MERGERS

So far, we have considered exogenous mergers. In this section, we review the literature on endogenous mergers, which is conceptually very different from the literature on exogenous



mergers. As stated in Nocke (2008, p. 578), “the term ‘endogenous mergers’ reflects the view in economic theory that mergers are equilibrium outcomes”. So this literature analyzes the firms’ incentives to merge and predicts the level of market concentration and the type of mergers that will arise in equilibrium.

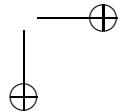
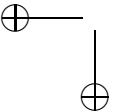
Focusing on the “market power” motive for mergers, one of the main questions addressed in the literature on endogenous mergers has to do with the limits to monopolization by acquisition and the relationship between concentration and market structure. A group of papers within this literature model the merger process as a bidding game or a non-cooperative coalition formation game. Among these papers, we can cite Kamien and Zang (1990), Gowrisankaran (1999), Pesendorfer (2005), Nocke (2000) and Vasconcelos (2006).

Kamien and Zang (1990) propose a formal three-stage non-cooperative game where owners benefit both from selling and buying firms and from operating them. Assuming constant marginal costs, in the first stage each owner bids for every other firm and announces an asking price at which she would sell the firm.² Once all bids and asking prices are known, firms are allocated to owners at prices equal to the new owner’s bid, which must be the maximal bid and above the original owner’s asking price. In the second stage, each owner decides how many firms to operate. In the third stage, active firms decide on output levels. It is shown that in order for an equilibrium with complete monopolization of the industry to exist, the industry should be initially sufficiently concentrated. In industries with a large number of firms, only unmerged equilibria or partial monopolization equilibria may exist. Regarding the former, the authors show that an equilibrium with no merger always exists for sufficiently low bids (e.g. bids below the single firm profit in the initial n -firm oligopoly) and sufficiently high asking prices (e.g. above the monopoly profit). With respect to partial monopolization equilibria, it is shown that in such equilibria, one owner must possess more than 50 percent of all the firms in the industry. So the authors conclude that if antitrust authorities forbid any single owner from acquiring more than half of the industry’s total number of firms, even partial monopolization via acquisition could be prevented.

The intuition to understand the results in Kamien and Zang (1990) is the following: consider a candidate equilibrium in which one owner acquires the rest of the firms ($n - 1$) in the industry. Notice that in this setting, buying firms is expensive because, by not accepting a bid, a firm free-rides on the reduction in competition induced by the remaining acquisitions. In particular, the buyer must be willing to pay to each of these firms the duopoly profit because, otherwise, any of these firms would have an incentive to deviate from the candidate equilibrium in the first stage of the game by increasing its asking price above the highest bid received. In this case, the buyer would own only $(n - 1)$ firms and the industry would become a duopoly. So in order to prevent such deviation, the buyer has to pay each of the $(n - 1)$ sellers at least the duopoly profit. Of course, for a sufficiently large number of firms, the total price to be paid is higher than the monopoly profit, which is the upper bound of the buyer’s willingness to pay, and then no merger will occur.³

² The bids and asking prices are assumed to be all posted simultaneously, which seems to represent situations where there are no negotiations among buyers and sellers and where trade takes place through some form of trading mechanism, like a stock exchange.

³ Whereas Kamien and Zang (1990) assume that all firms have identical constant marginal costs of production, Kamien and Zang (1991) extend the results to the case of an industry with a convex cost function, and Kamien and Zang (1993) further extend the results by analyzing sequential acquisitions.



16 *Handbook of game theory and industrial organization: applications*

There are several limitations in the paper by Kamien and Zang (1990). One limitation is that there is a multiplicity of equilibria, a characteristic common to most of the papers on endogenous mergers. Another limitation is that the authors use a static model where firms take only short-term considerations into account. This may lead to the wrong conclusions because several dynamic determinants of firm behavior like entry, exit or investments should have important effects on the type of mergers that occur in equilibrium and on their impact on consumers and welfare.

Trying to build on the second limitation, Gowrisankaran (1999) develops a dynamic, infinite-horizon, discrete-time model of endogenous mergers where entry, exit as well as investment are taken explicitly into account and where each firm chooses the strategy that maximizes the expected discounted value of its profits. In particular, at each period of the game, first the firms can merge, which allows them to join their production capacities. Then an exit stage takes place, after which there is a production stage, then an investment stage that allows the firms to increase their capacity, and, to conclude, there is the possibility of entry. This one-period game is repeated an infinite number of times. The equilibrium concept used is the Markov perfect Nash equilibrium. In order to mitigate the problem of multiplicity of equilibria, which is common to all the models of endogenous mergers with multiple firms, the author proposes a sequential merger process where the largest firm has the possibility to acquire a smaller firm by paying its asking price. If the largest firm refuses to acquire a firm, then it is the second largest firm that can decide to acquire a smaller firm. If, on the other hand, the largest firm does acquire a smaller firm, the industry structure changes and the process starts again, that is, the largest firm in the new industry can acquire a smaller firm and so on. Every period, the merger process continues until there is only one firm left or until the second smallest firm decides not to buy the smallest one.

As expected, this complex game cannot be solved analytically, and requires the use of numerical methods to be able to compute the equilibrium. The author obtains that although monopoly never occurs in equilibrium, allowing for mergers increases industry concentration. Moreover, mergers serve as a quick way for the industry to adjust when needed, compared to exit. Somehow, mergers are a substitute for exit for firms that would otherwise exit the industry and have the possibility to be bought by other firms. The entry rate increases when mergers are allowed because even if entry is not profitable for a firm alone, it can reduce the profits of incumbents so much that the latter want to buy them. Also investment rates drop when mergers are allowed because a merger is also a substitute for investment in order to increase capacity.⁴

In the same vein, Pesendorfer (2005) studies a model that builds on the static model by Kamien and Zang (1990), analyzing endogenous mergers in a dynamic game with infinite (discrete) periods of time, $t = 1, 2 \dots \infty$, with gradual entry of firms. Every period, the following stage game is played. First, there is the possibility of entry; second, there is a merger game where active firms simultaneously bid to buy other firms and then, after observing all the offers, announce an asking price. Then, buyers and sellers are matched and, finally, all active remaining firms collect profits. The author obtains the Markov perfect Nash equilibria

⁴ In Gowrisankaran and Holmes (2004) a dynamic, dominant-firm model with endogenous mergers and output-investment decisions is developed to study the determinants of the evolution of industry concentration. The authors show that the answer depends on the initial industry concentration and on the elasticities of supply and demand and the discount factor.

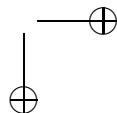
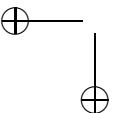
of this dynamic game, showing that firms are willing to merge whenever a current merger leads to additional mergers in the future, which increases future expected profits. So there is a “merger motive” for mergers in this dynamic setting that differs from the efficiency-improved argument used in the literature. In particular, it is shown that in a Cournot setting with linear demand and constant marginal costs, mergers that do not completely monopolize the industry can be profitable. Also, and in contrast to Salant et al. (1983), mergers that are not profitable when there are a small number of active firms, can become profitable when the number of them increases.

Concerning the welfare consequences of mergers and in contrast to Levin (1990) and Farrell and Shapiro (1990), it is shown that there can be profitable mergers in non-concentrated Cournot industries that reduce welfare. The intuition is that in this dynamic setting, there can be profitable mergers that yield losses in the short run but gains in the long run.

Nocke (2000) proposes a coalition formation game with costly entry. He considers two different models, an *exogenous sunk costs* model, where the only sunk costs are the set-up costs and an *endogenous sunk costs* model, where there is possibility for the firms to invest in R&D to increase consumers’ willingness to pay for the product (see Sutton, 1991, 1998). In both of them, the firms compete in prices with differentiated products. The timing of the game is as follows: first, firms decide whether or not to enter, paying an entry fee; second, the firms that enter form coalitions; third (only in the endogenous sunk costs model) the firms can invest a fixed amount to increase the quality of their products; and finally, the formed coalitions compete in prices. The author finds first that in exogenous sunk costs industries the upper bound to concentration goes to zero as market size goes to infinity. The intuition is that in a free entry equilibrium, an increase in the market size increases entry, given that entry profitability increases also. Then, it is not possible to sustain concentrated outcomes in large markets. In contrast, in endogenous sunk costs industries, the equilibrium number of entering firms remains finite no matter the size of the market, so the upper bound to concentration does not decrease with the size of the market.

On the same lines, Vasconcelos (2006) also studies the relationship between market size and concentration in exogenous and endogenous sunk costs industries, developing an endogenous coalition formation game where there is entry in the first stage, coalition formation in the second stage, R&D investment in the third stage and Cournot competition in the last stage. As in Nocke (2000), he shows that, whereas in exogenous sunk costs industries the upper bound to concentration falls as the market size increases, in endogenous sunk costs industries, regardless of the size of the market, arbitrarily concentrated outcomes can arise in equilibrium.

Most of the papers on endogenous mergers have focused on the market power motive for mergers. There are, however, several other motives for mergers that have been addressed in the literature. For example, Zhou (2008) analyzes endogenous horizontal mergers under cost uncertainty. In the first stage of the game, the firms vote on a merger proposal that would join those who accept it into a merged entity. Rejecting the proposal implies remaining independent. In the second stage, the firms compete in quantities. It is assumed that firms do not know what their future production costs will be at the time of the vote and that they privately learn their costs before production takes place. The advantage of the merged firm is that it allocates its production optimally among its facilities, reducing the expected costs. This advantage is balanced in equilibrium with the benefit of free-riding on the reduced competition brought about by the merger, when remaining independent. In contrast with



the general message that emerges from the literature that there is a tendency for firms to remain independent, Zhou (2008) finds that due to the positive effect of the merger, which increases with the level of uncertainty, mergers occur in this setting if and only if uncertainty is sufficiently large and also that more firms join the merger as the level of uncertainty increases.

Another motive for horizontal mergers that has been addressed in the literature is a preemptive merger taking place to avoid other alternative mergers to occur, which could lead the initial merging firm to end up being worse off than in the initial merger. For example, Fridolfsson and Stennek (2005) use preemption as the main argument to explain the empirical puzzle that mergers often reduce profits but increase the market value of the merging parties. They propose a model with a three-firm oligopoly where the merger game is modeled as a coalitional bargaining game with an infinite horizon where sequentially a randomly selected firm bids for other firms that can accept or reject the offer. The idea is that if becoming an insider is better than being an outsider, firms can rationally merge in order to prevent their partners from merging with the other rival firm. Even if the merger is unprofitable compared with the pre-merger situation, it may still be profitable compared with the alternative of another merger. The authors argue that if the stock market is efficient in the sense that share prices reflect actual firms' values, a (low) value of a firm pre-merger may reflect the risk that the firm becomes an outsider whereas, after the merger is announced, its value may increase because that firm will not become an outsider.

In the same vein, Brito (2003) proposes a two-stage model where there is a merger game in the first stage and spatial price competition (*à la Salop*) with differentiated products in the second stage. Brito shows that even when some outsiders gain more (and others less) than the insiders, the latter firms may still be interested in being insiders to avoid the possibility of being in the place of the least benefitted outsiders. Two crucial assumptions to obtain the result are first some degree of asymmetry in the rival's payoffs, which is attained with the spatial competition model because in this setting the effect of a merger on different firms depends on the location of those firms. Second, there must be a limited number of possible mergers imposed by the antitrust authority. So these types of mergers are more likely to occur in concentrated markets with firms selling non-symmetrically differentiated products.

Up to now we have focused on papers that model merger formation as a non-cooperative coalition formation game. An alternative approach is treating merger formation as a cooperative game. This approach does not fully specify the merger process but looks at whether a particular market structure can be an equilibrium outcome of a merger process because no firm (or group of firms) have an incentive to deviate and change the current configuration: in other words, the "stability" concept plays a central role in this branch of the literature.

For example, Horn and Persson (2001b) propose a game with two stages: in the first stage, owners form firms. A firm needs one unit of an asset to be able to produce. This asset is indivisible and each unit belongs to a separate owner, so each separate owner can run just one firm. But owners can also merge so that they can create firms that control more than one unit of the asset. In the second stage, the firms formed in the first stage compete non-cooperatively in an oligopolistic market. The authors show that the free-riding problem arising in most of the papers on mergers is not so pervasive in this setting under general assumptions on costs and demand. In particular, monopoly is shown to be the equilibrium outcome when complete monopolization is allowed. And when it is not, conditions are shown under which the most concentrated outcome allowed arises in equilibrium. In particular, no cost savings

are needed in order for mergers to less concentrated market structures than monopoly to take place in equilibrium.

Banal-Estañol, Macho-Stadler and Seldeslachts (2008) focus on the interactions between mergers and managers' investment decisions and on how the internal organization of the firms influences these interactions. They also use a cooperative game perspective, studying a game of endogenous mergers with three managers. Each manager controls some non-transferable resources and chooses with whom to merge. The resources of a newly formed firm are the sum of the resources that the participating managers control, which allows the authors to introduce efficiency gains of mergers in the model. But the efficiency gains can only be realized if the managers invest taking into account that investment is costly because there can be an alternative more market-oriented use of the resources; moreover, in absence of trust between the new firm's managers, each manager could prefer to free-ride on the other managers' investments. In this setting the authors show that when there is trust, managers invest more in a merged firm because of the existence of synergies. On the other hand, managers mainly choose the monopolistic industry structure. When there is no trust, however, the internal conflict can dominate the synergies and then larger firms tend to invest less than smaller ones, which implies that now, all industry structures can be stable outcomes. So, to conclude, the existence of an internal conflict leads to less concentration and when there are mergers in equilibrium, the merged firm could end up being less efficient compared with firms that would not merge.

4 HORIZONTAL MERGERS IN VERTICALLY RELATED INDUSTRIES

In this section, we deal with the analysis of horizontal mergers in vertically related industries, where firms are located at different stages of production or distribution. These types of industries have recently attracted considerable attention from policymakers, antitrust authorities and economists. In this setting, a horizontal merger, apart from reducing the level of competition at a particular level of the industry can also affect the terms of vertical trade between upstream and downstream firms, which determines the efficiency of the downstream firms and, ultimately, the final prices paid by consumers. For example, a horizontal merger downstream not only reduces downstream competition but could allow the merged firm to improve its bargaining position and to get better terms of trade when negotiating with the suppliers. The question would be then, under what conditions are these lower intermediate prices passed on to consumers so that, overall, the merger could be welfare enhancing?

The formal analysis of vertically related industries is complex. This is why many of the papers analyzing this topic consider either the case of an upstream (or downstream) monopoly or when there is competition at both levels of the industry, they consider two manufacturers and two retailers locked in exclusive relations. These simplifications reduce the complexity of these types of models but impose some limitations, as in many cases we find industries characterized by multiple interlocking relations in which the upstream firms compete with the same set of downstream firms.⁵

⁵ Nocke and Rey (2014) analyze a model of a vertically related industry with interlocking bilateral relations between upstream manufacturers and downstream retailers. They do not focus on horizontal but on vertical mergers, however, which are outside the scope of this chapter.

20 *Handbook of game theory and industrial organization: applications*

Among the papers analyzing horizontal mergers in vertically related industries we can distinguish between those analyzing upstream mergers from those analyzing downstream mergers. Within the former group, Horn and Wolinsky (1988), Ziss (1995), Milliou and Petrakis (2007) and Milliou and Pavlou (2013) study upstream mergers in models in which two manufacturers and two retailers are locked in exclusive relations by pairs, whereas in Inderst and Wey (2003) and Milliou and Sandonís (2014) this assumption is relaxed.⁶

Horn and Wolinsky (1988) analyze the profitability of an upstream merger in a setting with the upstream and downstream firms bargaining over a linear wholesale price. In contrast with what happens when the suppliers have the power to set prices, it is not necessarily the case that suppliers' profits are higher under monopoly than under duopoly due to a "bargaining effect" of the merger. The difference of behavior between a monopolist upstream and an independent upstream firm is that the former takes into account the cross-effect of the price of one downstream firm's input on the rival downstream firm's input demand. This is in favor of an upstream monopolist when it has the power to set prices; when there is bargaining with the downstream firm, however, if a lower price to one downstream firm increases the other downstream firm's input demand (which occurs when the goods are complements), this weakens the bargaining position of the upstream monopolist compared with that of an independent firm and leads to lower profits in the upstream industry under monopoly than with two independent suppliers. When downstream goods are substitutes, however, the bargaining effect would be positive and an upstream merger would be profitable. A similar reasoning would apply to downstream mergers.

Ziss (1995) assumes that the manufacturers produce differentiated final goods, sold via observable two-part tariff contracts to downstream independent retailers. In contrast with Horn and Wolinsky (1988), it is shown that under very general demand and cost conditions, an upstream merger that results in a multi-product monopolist is always profitable in this setting and more than that, it allows the merged firm to achieve the vertically integrated multi-product monopoly profits (which cannot be achieved when the manufacturers remain independent or even when they merge but use linear wholesale prices).⁷ As a result, upstream mergers in this setting are always profitable and anticompetitive.

Milliou and Petrakis (2007) study how the use of different types of contracts (linear vs two-part tariff contracts) and the possibility of bargaining, affect the incentives of upstream firms to merge and the welfare effects of these mergers. They use a setting with two manufacturers and two retailers locked in exclusive relations, where the manufacturers decide first whether or not to merge and the type of contract, then they bargain over the contract terms with retailers and, finally, there is competition downstream. It is shown that in the absence of any efficiency gains, the upstream firms have a disincentive to merge when they trade using two-part tariff contracts, which is a surprising result. Concerning welfare, it is shown that under endogenous contract choice, an upstream merger could be welfare enhancing whenever the merged firm uses a two-part tariff contract. However, this is never the case: when contract choice is endogenous, the merged firm never trades exclusively through two-part tariffs. This

⁶ Froeb, Tschantz and Werden (2007) analyze a setting with a monopoly downstream and two upstream firms producing differentiated goods and discusses how the effect of an upstream merger on the downstream market completely depends on the relationship between the merging manufacturers and the retailer.

⁷ O'Brien and Shaffer (1992) show that this result does not hold under secret two-part tariff contracts. A well-known opportunistic problem arises under secret contracts that prevents the upstream monopolist from achieving the full monopoly profits.

implies that in the absence of efficiency gains all upstream mergers should be forbidden. Interestingly, the anticompetitive effect of a merger does not arise from the increase in market concentration but from the contract distortion that the merger brings about.

Milliou and Pavlou (2013) use a similar setting to Milliou and Petrkis (2007) and analyze the incentives for upstream mergers in a setting in which efficiency gains produced by R&D at the upstream level are considered. It is shown that as long as downstream competition is not too strong, upstream mergers increase R&D investments, reduce the wholesale prices and the efficiency gain is passed on to consumers, who benefit from the merger.

Inderst and Wey (2003) also analyze a model of input price determination in a bilateral oligopolistic industry and analyze the incentives for mergers at the two levels. Nevertheless, in contrast with Horn and Wolinsky (1988), they do not require that each buyer is locked-in with a specific supplier and they consider efficient bargaining over non-linear pricing. They show that the upstream firms have an incentive to merge if the inputs they supply are substitutes. If they are complements, however, the merger is not profitable. Concerning downstream mergers, they are profitable only if unit costs in the upstream industry are strictly increasing. The authors then introduce a non-contractible technology choice upstream at the first stage and show that the incentives to adopt a cost-reducing technology are higher if upstream firms remain separated and downstream firms merge and also that downstream firms may strategically merge to affect the upstream technology choice, which may benefit all market participants, including consumers.

Milliou and Sandonís (2014) introduce product variety issues in this literature. They show that when the manufacturers distribute their products through multi-product retailers, an upstream merger, although it leads to an increase in the wholesale prices, can also incentivize the introduction of new varieties into the market, which is valuable for consumers. The product variety efficiency though, arises only when vertical relations are present: when manufacturers sell their products directly to consumers, a merger never results in more product variety. Still, both with or without vertical relations, an upstream merger is shown to be harmful to consumers and welfare.

Another branch of the literature on horizontal mergers in vertically separated industries has analyzed horizontal mergers in the downstream industry. A general result found in this literature is that an increase in the countervailing power of downstream firms (produced, for example, by a downstream merger) reduces the intermediate prices charged by suppliers, although the welfare effects are less clear. For example, when downstream firms bargain with a single supplier over linear tariffs, Von Ungern-Stenberg (1996), Dobson and Waterson (1997) and Chen (2003) show that downstream mergers can be welfare enhancing when there is strong enough competition downstream. Chipy and Snyder (1999) use a similar setting but consider non-linear contracts, and relate the results to the curvature of the profit function of the supplier. Lommerud, Straume and Sorgard (2005, 2006) consider a duopoly upstream and each downstream firm is assumed to be locked into a bilateral monopoly situation with its own independent input supplier. Suppliers set linear wholesale prices. They focus mainly on the profitability of downstream mergers in the first paper and on the comparison between national vs international mergers in the second. Symeonidis (2010) investigates how sensitive the sign of the welfare effect of downstream mergers is to the mode of downstream competition (prices vs quantities) and on whether they bargain over linear or two-part tariff contracts.

In Faulí-Oller and Sandonís (2016), the welfare consequences of downstream mergers are related to the market structure of the upstream sector in a setting with one dominant

upstream firm offering observable two-part tariff contracts, and downstream firms competing à la Cournot and an (exogenous) alternative supplier. It is shown that a merger downstream leads to lower wholesale prices and when the upstream industry is sufficiently concentrated these lower intermediate prices are translated to final consumers in the form of a lower final price that benefits consumers and social welfare. This result supports the view that symmetry between upstream and downstream markets increases welfare (Inderst and Shaffer, 2008).⁸

There are two papers that consider product variety issues within the literature on downstream horizontal mergers, namely, Inderst and Shaffer (2007) and Faulí-Oller (2008). Both papers show that a merger among retailers allows them to commit not to sell one of the goods supplied by manufacturers, and thus, that such a merger can result in a welfare-detrimental decrease in product variety.

Faulí-Oller, Sandonís and Santamaría (2011) show that downstream mergers increase the incentives of an upstream firm to invest in cost-reducing R&D. The upstream firm revenues increase with total industry profits, which in turn increase with downstream concentration, and this explains the positive link between concentration and investment. This effect is so important that it outweighs the negative effect of the merger on prices due to lower competition. Therefore, in their context, horizontal mergers are pro-competitive.

5 INTERNATIONAL SETTING

The study of horizontal mergers in an international context deserves a specific treatment because the papers dealing with this issue share special features. The conventional wisdom is that economic integration promotes both domestic and cross-border mergers. Papers usually measure integration by the additional cost of exporting in comparison to national production. This additional cost (t) includes both tariffs and transportation costs. A reduction of t means that economic integration increases.

Many papers study how the profitability of mergers changes with integration. Long and Vousden (1995) obtain that a marginal reduction in t (in the neighborhood of free trade) increases the profitability of domestic mergers if merging firms are asymmetric enough. Other papers have studied non-marginal changes in t . Gaudet and Kanouni (2004) show that free trade stimulates domestic mergers if the initial tariff is over the prohibitive level. This corresponds with the fact that the Canada–United States Free Trade Agreement has stimulated domestic mergers in Canada (Breinlich, 2008). Bertrand and Zitouna (2006) find that the relationship between profitability of a cross-border merger and t has an inverted U-shape (for a related result see Benchekroun and Chaudhuri, 2006).

Bjorvatn (2004) allows two foreign firms to serve a national market in three different ways: acquisition of a unique national firm, greenfield investment or exports. Reductions in t increase the profitability of the cross-border merger when the outsider switches from serving the market through a greenfield investment to serving it through exports. This reduces competition in the market and increases the profitability of the merger.

Horn and Persson (2001a) consider a case with four firms. Firms 1 and 2 are located in one country, whereas firms 3 and 4 are located in a second country. Monopolization is forbidden

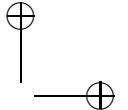
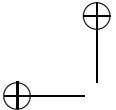
⁸ The same result is found in Faulí-Oller and Bru (2008) using a similar setting but under the assumption of secret contracts.

and there are no costs apart from t . They study the equilibrium ownership structure using the cooperative concept developed in Horn and Persson (2001b). They obtain that for low values of t two international mergers (MI) arise in equilibrium, whereas for high values of t we have two national mergers (MN). This last result seems surprising, because the higher is t the higher the cost reduction obtained through an international merger. The reason for the result is the following: when comparing MI with MN we have that all firms are decisive. Therefore, the structure that dominates is the one that maximizes industry profits. In MN in each market we have a firm with no costs and a firm with cost t . In MI in each market we have two firms with zero costs. An increase in t has two effects on industry profits: one is positive and the other is negative. The latter is because as t increases, costs also increase (efficiency effect) and the former is because as t increases competition decreases (competition effect). We have that the competition effect dominates for high values of t and this explains the result. Summarizing, Horn and Persson (2001a) find that reductions in t may change the market structure from national mergers to international (cross-border) mergers.

Chaudhuri (2014) obtains that cross-border mergers are more likely to occur when markets are segmented. He considers two symmetric segmented markets (markets A and B). In each market, there is a national firm that only sells in its own market. There is also a multinational firm that sells in both markets. Costs are strictly convex. This implies that the two markets are interrelated: what the multinational enterprise (MNE) sells in market A affects its marginal cost and therefore it affects what it wants to sell in market B. This interrelationship determines the effect of a merger. If the MNE buys the firm in A, it will contract the output in A. This will reduce its marginal costs, which provides incentives to sell more in B. Therefore, the effect of buying the firm in market A is that the price increases in A and decreases in B. If markets were integrated, the merger with the firm in A would increase the price in both markets. This effect is very important because it reduces the profits to be obtained while rejecting a merger offer. Suppose that a multinational makes offers to buy both firms. Then, the outside option of national firms is lower in segmented markets because if it rejects the offer the price will decrease by more if markets are segmented than if they are integrated. The authors claim that international mergers have concentrated in industries with segmented markets, like services.

Neary (2007) studies cross-border mergers in a general equilibrium setting. This has the advantage that the costs of firms are endogenously determined. We have two countries, a home and a foreign country. Suppose that we start with a situation in which there is free trade but cross-border mergers are not allowed. Then, each country specializes according to its comparative advantage, which means that the cost of the different sectors will vary between countries. For some sectors, costs will be higher in the home country and, for other sectors, costs will be higher in the foreign country. The paper analyzes the consequences of allowing for cross-border mergers. It is found that, in a Cournot setting, cost differentials generate incentives to merge and that previous mergers stimulate future mergers. Therefore, the author finds that a wave of cross-border mergers arise where low-cost firms buy high-cost firms. As a result, mergers reinforce the specialization of countries according to their comparative advantage.

Nocke and Yeaple (2007) also study cross-border mergers in a general equilibrium setting. The objective of cross-border mergers in this paper is not gaining market power but acquiring capabilities from target firms. The authors consider that firms have two types of capabilities: mobile and non-mobile. Mobile capabilities can be applied equally well to either national or foreign markets. Non-mobile capabilities instead are country specific in the sense that



24 *Handbook of game theory and industrial organization: applications*

local firms have an advantage over foreign firms to operate the local market. To clarify the difference, it is useful to think of mobile capabilities as technological capabilities and non-mobile capabilities as marketing capabilities. Cross-border mergers allow foreign firms to acquire these better marketing capabilities than local firms have. This acquisition of marketing capabilities is the critical difference between cross-border mergers and greenfield foreign direct investment. The paper's main result is to establish the characteristics of firms that engage in cross-border mergers. They depend on the source of firm heterogeneity. If firms differ basically in their mobile capabilities, the more efficient firms are the ones engaging in cross-border mergers. If instead firms differ in their non-mobile characteristics, the cross-border mergers are carried out by the least efficient firms.

6 WELFARE

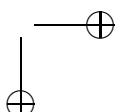
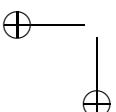
The 2010 US Horizontal Merger Guidelines state: “The higher the post-merger HHI and the increase in the HHI, the greater are the Agencies’ potential competitive concerns and the greater is the likelihood that the Agencies will request additional information to conduct their analysis”. (HHI stands for the Herfindahl-Hirschman Index, which measures the sum of the squares of all the firms’ market shares. The post-merger HHI is calculated assuming that the post-merger market shares of outsiders are equal to their pre-merger level and the post-merger market share of insiders is equal to the sum of their pre-merger market shares. For example, the increase in the HHI index due to the merger between firm i and firm j is $2s_i s_j$ where s_j stands for pre-merger market shares.) We will check whether both ideas are corroborated in the following models.

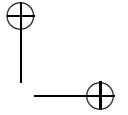
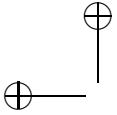
6.1 Static Setting

Farrell and Shapiro (1990) analyze mergers in a Cournot setting with homogeneous goods. $P(Q)$ denotes market demand and $C_i(q_i)$ each firm’s cost function. Demand is downward sloping and satisfies $P'(Q) + q_i P''(Q) < 0$. Costs satisfy $C_i''(q_i) > P'(Q)$. Denoting by $r_i(q_{-i})$ the best-response function of firm i , we have that $(-1) < r'_i(q_{-i}) < 0$ and the Cournot equilibrium is stable. They derive conditions for a merger to increase either consumer surplus or total welfare. As far as consumer surplus is concerned, they obtain that a merger that does not generate synergies (defined in Section 2.1.2) increases the price. Furthermore, they add that “huge” efficiencies, in terms of reduction of costs, are needed for a merger to reduce price.

As far as total welfare is concerned, they recognize that one of the problems to evaluate the effect of mergers on total welfare is that the antitrust authority does not know the evolution of the costs of merging firms (insiders) after the merger. To side-step this problem, they propose to study the external effect of mergers, i.e. their effect on consumers and non-participating firms (outsiders). As proposed mergers are likely to be profitable, a positive external effect will be a sufficient condition for a merger to increase total welfare.

A merger implies a reduction in the number of independent firms. But Farrell and Shapiro (1990) look at the situation differently. We know that merging firms (unless costs efficiencies are very high) will reduce their output after the merger. So a merger is studied as a differentiable process where the merging firms reduce their output from the pre-merger level





to the post-merger level. First of all, let us study the effect of a marginal reduction in the output of the merging firms on the external effect, what they call an infinitesimal merger. Let λ_i measure how firm i optimally adjusts its production when total output increases. It can be calculated using the expression of the slope of the reaction function:

$$-\lambda_i = \frac{r'_i(q_{-i})}{1 + r'_i(q_{-i})} < 0.$$

The infinitesimal effect of a merger amounts to:

$$dEW = \left(\sum_{i \in O} \lambda_i s_i - s_I \right) P'(Q) Q dQ,$$

where I refers to insiders and O to outsiders. The external effect of an infinitesimal reduction on the output of insiders is positive if:

$$\eta \equiv \sum_{i \in O} \lambda_i s_i - s_I \geq 0,$$

given that $dQ < 0$, because a merger reduces total output.

We are interested in calculating the overall effect of a merger. Then, we calculate it by adding the infinitesimal effects. $Q_I^{initial}$ is the (initial) output of the merging firms pre-merger and Q_I^{final} the (final) output of the merging firms post-merger. We have that $Q_I^{initial} > Q_I^{final}$. So the external effect of the merger can be written as:

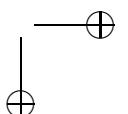
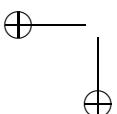
$$\Delta EW = \int_{Q_I^{initial}}^{Q_I^{final}} \eta(Q_I) P'(Q) Q \frac{dQ}{dQ_I} dQ_I,$$

where for each Q_I , the integrand is evaluated assuming a Cournot equilibrium among outsiders, given Q_I . Given that we are considering output reducing mergers ($dQ_I < 0$), if $\eta(Q_I) \geq 0$ the previous expression is positive. A sufficient condition for this is that

$$\eta(Q_I^{initial}) = \sum_{i \in O} \lambda_i s_i^{initial} - s_I^{initial} \geq 0 \quad (2.2)$$

and $\eta'(Q_I) \leq 0$. The latter condition holds if $P'' \geq 0, P''' \geq 0, C_i'' \geq 0, C_i''' \leq 0$.

Assuming that the previous conditions hold, the external effect is positive if (2.2) holds. It will hold if the merging firms are small. It coincides with the spirit of the Horizontal Merger Guidelines that approve mergers when they slightly increase the HHI. There are two additional elements that make it more likely that the external effect is positive: the size of outsiders and their reaction. The greater the size of outsiders the more likely that the external effect is positive because their output expansion after the merger has a greater positive effect on welfare because they have higher price-cost margins. It is precisely the need for big outsiders that explains that one should only approve the merger of small firms. The greater the reaction of outsiders the more likely it is that the external effect is positive. Observe that if the reaction



26 Handbook of game theory and industrial organization: applications

was so great that price remains constant after the merger, the external effect would be positive, because consumer surplus would not change and profits of outsiders would increase (observe that after the merger they still maximize and the output of the other firms is lower because they produce more). On the other hand, if outsiders “did not respond, that is if $\lambda_i = 0$ for $i \in O$, then every output reduction would be bad for rivals and consumers jointly: rivals would benefit, but consumers would lose by more” (Farrell and Shapiro, 1990, p. 115).

In the case of linear costs and demand and for symmetric firms (Salant et al., 1983), we have that $\lambda_i = 1$. Then (2.2) reads $\sum_{i \in O} s_i^{initial} - s_I^{initial} \geq 0$, i.e. the market share of outsiders is greater than that of insiders, i.e. the market share of merging firms is lower than 50 percent.⁹ In the case of linear demand and quadratic costs (Perry and Porter, 1985 and McAfee and Williams, 1988), we have that $\lambda_i = \frac{s_i}{\varepsilon}$, where ε refers to the elasticity of demand. Then the condition is $s_I^{initial} < \frac{1}{\varepsilon} \sum_{i \in O} (s_i^{initial})^2$. In this case, it is good to have big outsiders, not only because their output expansion has a greater positive impact on welfare, but also because they simply react more. We have that the greater the concentration the more likely that the external effect is positive. This goes against the presumption in the Horizontal Merger Guidelines that the higher the concentration the more likely that a merger raises potential competitive concerns.

The analysis in Farrell and Shapiro (1990) ignores the possibility of entry. If entry does occur, the effect of a given merger will be very different.¹⁰ Anderson, Erkal and Piccinin (2015) analyze aggregative non-cooperative games with endogenous entry (aggregative games are games where the players’ payoff can be expressed as a function of their own action and an aggregate of all players’ actions) and apply the analysis to different settings, including mergers. They show neutrality properties of mergers in the long run, in the sense that entry just undoes the short-run effect of the merger, in that the aggregate remains the same, as well as outsiders’ actions and profits. Insiders’ profits, however, are weakly lower, which suggests that efficiency gains are needed in order for mergers to be profitable for the merging firms in the long run. Concerning consumer surplus, it is shown that it remains constant in the long run after a merger, the reason being that the positive effect of entry through more variety exactly compensates for the anticompetitive effect of the merger. An interesting implication of this analysis is that under free entry, mergers are welfare enhancing if and only if they are profitable for the merging firms in the long run.

6.2 Dynamic Setting

Nilssen and Sorgard (1998) study two sequential mergers. They consider the consequences of approving the mergers by considering their effect on the external effect (Farrell and Shapiro, 1990) in isolation. When we have only one merger, a positive external effect guarantees that the merger increases welfare because if the merger is proposed it must be because it is profitable. But when merger decisions are interdependent, the external effect can be overestimated, because the merger in isolation may be unprofitable. However, it takes place

⁹ Levin (1990) also obtains the 50 percent threshold in a setting with constant but asymmetric marginal costs. The additional assumptions are concavity of demand and that the most efficient outsider is less efficient than the most efficient insider.

¹⁰ The literature finds that the likelihood of entry after merger is higher in models of spatial differentiation à la Salop (Cabral, 2003) than in logit demand models (Werden and Froeb, 1998).

because of its strategic effect on the merger decisions of other firms. The interdependence of merger decisions should be taken into account while designing the optimal merger policy. The authors show, for example, that a sequence of two mergers may reduce social welfare while every merger taken in isolation increases the external effect.

It is not difficult to understand that when more than one merger is involved, the merger policy can be very complicated. Therefore, the result in Nocke and Whinston (2010) is striking, where the optimal dynamic merger policy is myopic (the antitrust authority is assumed to maximize consumer surplus). They consider a dynamic merger model with T periods. Firms are divided into k disjoint sets and they compete à la Cournot with homogeneous goods and asymmetric marginal costs. The merger of each set becomes feasible in period t with some probability. Once they become feasible, the post-merger marginal cost is realized. The authors show that one can compute the optimal merger policy following this local analysis. In each period the antitrust authority should approve a proposed merger if it reduces the price given the current market structure. The key point to understand the result is complementarity of merger decisions. A merger will only reduce price if the merger reduces cost significantly. The importance of the cost reduction should be related to price. Given a cost reduction, it is more significant if the price is low. Then, they obtain that if a merger reduces prices in isolation it will also reduce prices when another price-reducing merger takes place. For the moment, assume that all feasible mergers are proposed. The antitrust authority may ignore the future: a price-reducing merger can only have a positive effect, because it may stimulate future price-reducing mergers. Price-increasing mergers must be forbidden, because of their direct effect on price and because they may turn price-decreasing mergers into price-increasing mergers. Finally, they show that proposing a merger is a weakly dominant strategy. If the merger is proposed and it is not approved, profits of involved firms are as if the merger was not proposed. If the merger is proposed and it is approved it means that it reduces the price. But price-reducing mergers are profitable because of the gains in efficiency and because the reduction of the output of competitors.

6.3 Welfare Standards and Merger Rules

There has been a long discussion on whether antitrust authorities should maximize either total welfare or consumer surplus. It seems in practice that antitrust authorities behave as if they were enforcing a consumer surplus standard. However, the literature on merger rules, which we review below, emphasizes that the objectives that antitrust authorities enforce may be different from the ultimate goal of the merger policy. The reason for this surprising result is that the antitrust authority interacts with firms that are both strategic and better informed about merger characteristics.

Besanko and Spulber (1993) consider a merger to monopoly that involves Williamson's trade-off (Williamson, 1968): on the one hand, it reduces competition but, on the other hand, it reduces the marginal cost of production. The lower the post-merger marginal cost, the higher the profits and the consumer surplus induced by the merger, because the post-merger price is lower. Therefore, there is an alignment between private and social incentives. The authors consider the case where the post-merger marginal cost is private information of the merging parties and the antitrust agency only knows its probability distribution. The order of moves is as follows. In the first stage, a social welfare-maximizing policymaker chooses α , which determines the objective function of the antitrust authority: $\alpha\Pi + (1 - \alpha)S$, where Π stands

for profits and S for consumer surplus. Observe that this objective formulation includes total welfare ($\alpha = \frac{1}{2}$) and consumer surplus ($\alpha = 0$) as particular cases. In the second stage, the firms decide whether to merge or not. If they decide to merge they pay a fixed cost. Finally, the antitrust authority decides whether to challenge the merger or not. In order to understand the result we can compare the probability of challenging a merger if the policymaker can commit to it ex ante (β^f) to the one we have in equilibrium if $\alpha = \frac{1}{2}$ (β^i). We have that $\beta^f > \beta^i$. Why is that? By increasing the challenging probability, only mergers with great cost savings are proposed given that there is a fixed cost on proposing a merger. But once this self-selection has taken place, the antitrust authority wants to challenge the merger with a lower probability. There is a time inconsistency problem, that can be solved by endowing the antitrust authority with an objective that gives more weight to consumer surplus than to profits ($\alpha < \frac{1}{2}$). The situation is akin to the one we have in monetary policy (Rogoff, 1985). “The president will want to select a central banker who places greater weight on controlling inflation than on unemployment as compared to the president” (Besanko and Spulber, 1993, p. 4).

Lyons (2003) provides the following illustration. He considers four firms located equidistantly along the Salop circle. At each point demand is elastic and, therefore, higher prices suppose lower total welfare. He considers that only two mergers are possible: a merger between two neighboring firms (say A and B) and another one between non-neighboring firms (say A and C). The merger between A and B increases prices, which reduces total welfare in $\Delta < 0$ and reduces fixed costs in F . The merger between A and C does not change prices and reduces fixed costs in $(1 - \lambda)F$. Merger A–B is more profitable than merger A–C. Suppose that $(1 - \lambda)F > F + \Delta > 0$. In this case, the merger A–C increases total welfare by more than the merger A–B. However, if the antitrust authority maximizes total welfare, the merger A–B will be proposed and will be approved. If the antitrust authority maximizes consumer surplus, merger A–B will be rejected and merger A–C, which has no effect on consumer surplus, will be accepted. So, in this case, an antitrust authority with a consumer surplus standard, achieves a greater total surplus because it forces firms to choose the (profitable) merger that increases total welfare the most (see also Fridolfsson, 2007).

In Neven and Röller (2005), in spite of the fact that the social objective is social welfare, there is a discussion on whether the antitrust authority should be endowed with a social welfare standard or a consumer surplus standard. Why can it be better for a consumer surplus standard to pursue social welfare maximization? The reason is the following. In their model, consumers are passive players but firms can behave as lobbyists that try to affect the decision of the antitrust authority regarding mergers. On some occasions, to counterweight the power of lobbies and give consumers some voice, the consumer surplus standard performs better than a social welfare standard. In particular, the consumer welfare standard works very well to reduce the cases where a welfare-reducing merger is approved (Type II error) although it increases the cases where a welfare-increasing merger is blocked (Type I error).

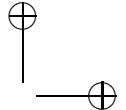
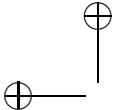
Armstrong and Vickers (2010) study a case where a principal delegates the choice of a project to an informed agent. The project is fully described by two scalars u and v : agent's payoff is u while the payoff of the principal is $v + \alpha u$, where $\alpha \geq 0$. Support of (u, v) is a rectangle $[0, U_{max}] \times [v_{min}, v_{max}]$. The agent knows the characteristics of all possible projects while the principal only knows the characteristics of the project that is chosen by the agent. The principal may affect the choice of the agent by defining a set of permitted projects (D) that is a subset of $[0, U_{max}] \times [v_{min}, v_{max}]$, i.e. it may establish the characteristics that must be satisfied by the chosen project.

The translation to a merger enforcement setting is straightforward. The principal would be the antitrust authority and the informed agent the firms that aim to merge. Parameter v would be consumer surplus and u the profits of the firms. The firms maximize profits and the antitrust authority maximizes consumer surplus plus α times the profits of firms. If $\alpha = 1$, the antitrust authority maximizes total welfare and if $\alpha = 0$ it maximizes consumer surplus.

It is shown that the choice of D can be reduced to choose a threshold rule $r()$ such that (u, v) belongs to D if and only if $v \geq r(u)$. As a benchmark, define the naive threshold rule $r_{naive}(u) = -\alpha u$, such that D includes all projects that increase the payoff of the principal. The main result of the paper is that the optimal threshold rule $r^*(u)$, satisfies $r^*(u) > r_{naive}(u)$, when $u > 0$. “Therefore, the principal forbids some strictly desirable projects (and never permits an undesirable project)” (pp. 223–224).

In Nocke and Whinston (2013), we have a Cournot setting with homogeneous goods and asymmetric constant marginal costs. The lower the marginal cost the higher the output produced by firms, i.e. big firms have low marginal costs. There is one firm that can decide to merge with one of the remaining firms. Mergers reduce the marginal cost of production. Firms know the cost reduction of any merger while the antitrust authority only knows the cost reduction of the merger that is proposed. The antitrust authority chooses the optimal merger policy in order to maximize consumer surplus. The optimal merger policy is designed to correct the misalignment between private and social incentives, which is the following: firms will choose the merger that increases profits the most while the antitrust authority wants them to choose the merger that increases consumer surplus the most, i.e. reduces price the most. Firms have a bias to merge with bigger firms, because, given a cost reduction, it is when the profits increase more. To correct this bias the antitrust authority imposes, to approve a merger, stricter rules in terms of costs reduction to mergers involving bigger firms. This corresponds to the practice of US Horizontal Merger Guidelines, which is more suspicious of mergers with high pre-merger market shares.

Burguet and Caminal (2015) assume there are three (1, 2 and 3) identical firms pre-merger. Two-firm mergers reduce marginal costs. Merger to monopoly is forbidden. When firms 1 and 2 merge, the marginal cost becomes c_{12} . With the other two possible two-firm mergers involving firm 3, marginal cost becomes c_3 . We have that $c_{12} < c_3$. These costs reductions are common knowledge for firms. The antitrust authority only knows their probability distribution and does not know either the identity of firms. The antitrust authority maximizes expected consumer surplus. If the post-merger marginal cost is lower than c_n , the merger reduces price. The antitrust authority commits to allow mergers if the post-merger cost is lower than \bar{c} . The main result of the paper concerns the relationship between c_n and \bar{c} . Unlike Nocke and Whinston (2013), there is no conflict of interest between firms and the antitrust authority: the most profitable merger is the one that reduces price the most. Mergers are decided according to a bargaining protocol that yields a unique equilibrium. The important point, however, is that less efficient mergers occur with positive probability. This is what it is called a bargaining failure. The antitrust authority can reduce the bargaining failure by making the less efficient merger illegal by lowering \bar{c} . This positive effect has to be balanced with the fact that by lowering \bar{c} some price-reducing mergers are blocked. This last effect is arbitrarily small when $\bar{c} = c_n$ and, therefore, we have that in equilibrium $\bar{c} < c_n$. As in Armstrong and Vickers (2010), we have that some price-reducing mergers are forbidden while price-increasing mergers are never allowed.



7 CONCLUSIONS

Mergers are a very important economic phenomenon that affect consumers, firms and markets. The effect of a merger on consumers is ambiguous and depends on whether the merger leads to efficiency gains that are passed on to consumers in the form of lower prices, higher quality or new products and services. In absence of an effective merger control by an active competition authority, however, mergers may lead to an excessive market concentration and/or anticompetitive behaviors that could result in higher prices paid by consumers, or in lower-quality goods.

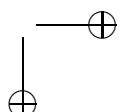
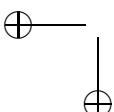
In this chapter, we have reviewed the theoretical literature that formally analyzes the profitability and the welfare consequences of horizontal mergers. Unfortunately, there is not a unique unified theoretical framework with which to analyze this topic. There are different types of mergers that require different theoretical treatments and different approaches have been used by researchers to explain the different relevant aspects about horizontal mergers.

In some papers, the effect of a given merger between an exogenous set of firms is studied (exogenous merger literature), whereas in others, firms' incentives to merge are explicitly analyzed and mergers are equilibrium outcomes (endogenous merger literature). Some papers only focus on short-term considerations (static models), whereas in others, long-term decisions, like entry, exit or investments are taken into consideration in the analysis (dynamic models). In some papers, merger formation is modeled as a bidding game or a non-cooperative coalition formation game, whereas in others, merger formation is treated as a cooperative game where the "stability" concept plays a central role.

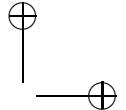
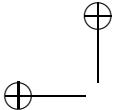
Undoubtedly, there is room for improvement. The literature on endogenous mergers would need more general and fully dynamic merger games to sharpen the results. At present, results depend too much on the structure of the merger game being analyzed. This explains the diverging results obtained in the literature. Related to this, the literature on mergers and welfare should further study the optimal merger policy in dynamic settings, extending and generalizing the results obtained in the pioneering paper Nocke and Whinston (2010). This would help to better understand the optimal behavior of the antitrust authority. The literature on cross-border mergers would improve if the analysis is embedded in a general equilibrium framework (as in Neary, 2007 or Nocke and Yeaple, 2007). This would allow a better understanding of the effect of mergers in the whole economy. These are just some examples of possible future research avenues in the field.

REFERENCES

- Anderson, S.P., Erkal, N. and Piccinin, D. (2015), "Aggregative oligopoly games with entry", mimeo.
- Andrade, G., Mitchell, M. and Stafford, E. (2001), "New evidence and perspectives on mergers", *Journal of Economic Perspectives* 15, 103–120.
- Armstrong, M. and Vickers, J. (2010), "A model of delegated project choice", *Econometrica* 78 (1), 213–244.
- Banal-Estañol, A., Macho-Stadler, I. and Seldeslachts, J. (2008), "Endogenous mergers and endogenous efficiency gains: the efficiency defence revisited", *International Journal of Industrial Organization* 26, 69–91.
- Benckokroun, H. and Chaudhuri, A.R. (2006), "Trade liberalization and the profitability of mergers. A global analysis", *Review of International Economics* 14, 941–957.
- Bertrand, O. and Zitouna, H. (2006), "Trade liberalization and industrial restructuring: the role of cross-border mergers and acquisitions", *Journal of Economics and Management Strategy* 15, 470–515.
- Besanko, D. and Spulber, D.F. (1993), "Contested mergers and equilibrium antitrust policy", *Journal of Law, Economics and Organization* 9 (1), 1–29.

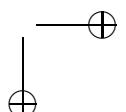
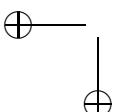


- Bjorvatn, K. (2004), "Economic integration and the profitability of cross-border mergers and acquisitions", *European Economic Review* 48, 1211–1226.
- Braid, R.M. (1986), "Stackelberg price leadership in spatial competition", *International Journal of Industrial Organization* 4, 439–449.
- Breinlich, H. (2008), "Trade liberalization and industrial restructuring through mergers and acquisitions", *Journal of International Economics* 76, 254–266.
- Brito, D. (2003), "Preemptive mergers under spatial competition", *International Journal of Industrial Organization* 21 (10), 1601–1622.
- Brito, D. and Catalao-Lopes, M. (2006), *Mergers and Acquisitions: The Industrial Organization Perspective*, Amsterdam: Kluwer Law International.
- Burguet, R. and Caminal, R. (2015), "Bargaining failures and merger policy", *International Economic Review* 56 (3), 1019–1041.
- Cabral, L. (2003), "Horizontal mergers with free-entry: why cost efficiencies may be a weak defense and asset sales a poor remedy", *International Journal of Industrial Organization* 21, 607–623.
- Chaudhuri, A.R. (2014), "Cross-border mergers and market segmentation", *Journal of Industrial Economics* 62, 229–257.
- Chen, Z. (2003), "Dominant retailers and the countervailing-power hypothesis", *RAND Journal of Economics* 34, 612–625.
- Cheung, F.K. (1992), "Two remarks on the equilibrium analysis of horizontal merger", *Economics Letters* 40, 119–123.
- Chipty, T. and Snyder, C. (1999), "The role of buyer size in bilateral bargaining: a study of the cable television industry", *Review of Economics and Statistics* 81 (2), 326–340.
- Compte, O., Jenny, F. and Rey, P. (2002) "Capacity constraints, mergers and collusion", *European Economic Review* 46, 1–29.
- d'Aspremont, C. and Jacquemin, A. (1988), "Cooperative and noncooperative R&D in duopoly with spillovers", *American Economic Review*, 78, 1133–1137.
- Deneckere, R. and Davidson, C. (1985), "Incentives to form coalitions with bertrand competition", *RAND Journal of Economics* 16, 473–486.
- Dobson, P.W and Waterson, M. (1997), "Countervailing power and consumer prices", *The Economic Journal* 107, 418–30.
- Dutz, M. (1989), "Horizontal mergers in declining industries: theory and evidence", *International Journal of Industrial Organization* 7, 11–33.
- Falvey, R.E. (1998), "Mergers in open economies", *The World Economy* 21, 1061–1076.
- Farrell, J. and Shapiro, C. (1990), "Horizontal merger: an equilibrium analysis", *American Economic Review* 80 (1), 107–126.
- Faulí-Oller, R. (1997), "On merger profitability in a Cournot setting", *Economics Letters* 54, 75–79.
- Faulí-Oller, R. (2000), "Takeover waves", *Journal of Economics and Management Strategy* 9, 189–210.
- Faulí-Oller, R. (2002), "Mergers between asymmetric firms: profitability and welfare", *The Manchester School* 70 (1), 77–87.
- Faulí-Oller, R. (2008), "Capacity restriction by retailers", WP-AD 2008-02, IVIE.
- Faulí-Oller, R. and Bru, L. (2008), "Horizontal mergers for buyer power", *Economics Bulletin* 12, 1–7.
- Faulí-Oller, R. and Motta, M. (1996), "Managerial incentives for takeovers", *Journal of Economics and Management Strategy* 5, 497–514.
- Faulí-Oller, R. and Sandón, J. (2016), "Welfare effects of downstream mergers and upstream market concentration", *The Singapore Economic Review*, forthcoming.
- Faulí-Oller, R., Sandón, J. and Santamaría, J. (2011), "Downstream mergers and upstream investment", *The Manchester School* 79, 884–898.
- Filson, D. and Songsamphant, B. (2005), "Horizontal mergers and exit in declining industries", *Applied Economics Letters* 12 (2), 129–132.
- Fridolfsson, S. (2007), "A consumer surplus defense in merger control", in V. Ghosal and J. Stenek (eds), *The Political Economy of Antitrust*, Amsterdam: Elsevier, 287–302.
- Fridolfsson, S. and Stenek, J. (2005), "Why mergers reduce profits and raise share prices: a theory of preemptive mergers", *Journal of the European Economic Association* 3 (5), 1083–1104.
- Froeb, L., Tschantz, S. and Werden, G. (2007), "Vertical restraints and the effects of upstream horizontal mergers", V. Ghosal and S. Stenek (eds), in *The Political Economy of Antitrust*, Amsterdam: Elsevier, 369–381.
- Fumagalli, E. and Vasconcelos, H. (2009), "Sequential cross-border mergers", *International Journal of Industrial Organization* 27, 175–187.
- Gaudet, G. and Kanouni, R. (2004), "Trade liberalization and the profitability of domestic mergers", *Review of International Economics* 12 (3), 353–358.
- González-Maestre, M. and López-Cuñat, J. (2001), "Delegation and mergers in oligopoly", *International Journal of Industrial Organization* 19 (8), 1263–1279.

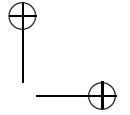
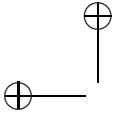


32 *Handbook of game theory and industrial organization: applications*

- Gowrisankaran, G. (1999), "A dynamic model of endogenous horizontal mergers", *RAND Journal of Economics* 30, 56–83.
- Gowrisankaran, G. and Holmes, T. (2004), "Mergers and the evolution of industry concentration: results from the dominant firm model", *RAND Journal of Economics* 35, 561–582.
- Horn, H. and Persson, L. (2001a), "The equilibrium ownership of an international oligopoly", *Journal of International Ecconomics* 53, 307–333.
- Horn, H. and Persson, L. (2001b) "Endogenous mergers in concentrated markets", *International Journal of Industrial Organization* 19, 1213–1244.
- Horn, H. and Wolinsky, A. (1988), "Bilateral monopolies and incentives for mergers", *RAND Journal of Economics* 19, 408–419.
- Inderst, R. and Shaffer, G. (2007), "Retail merger, buyer power and product variety", *Economic Journal* 117, 45–67.
- Inderst, R. and Shaffer, G. (2008), "Buyer power in merger control", in W.D. Collins (ed.), *ABA Antitrust Section Handbook, Issues in Competition Law and Policy*, Chicago, IL: American Bar Association, 1611–1636.
- Inderst, R. and Wey, C. (2003), "Bargaining, mergers and technology choice", *RAND Journal of Economics* 34 (1), 1–19.
- Kamien, M. and Zang, I. (1990), "The limits of monopolization through acquisition", *The Quarterly Journal of Economics* 105, 465–499.
- Kamien, M. and Zang, I. (1991), "Competitively cost advantageous mergers and monopolization", *Games and Economic Behavior* 3 (3), 323–338.
- Kamien, M. and Zang, I. (1993), "Monopolization by sequential acquisition", *Journal of Law, Economics and Organization* 9 (2), 205–229.
- Levin, D. (1990), "Horizontal mergers: The 50-percent benchmark", *American Economic Review* 80 (5), 1238–1245.
- Levy, D. and Reitzes, J. (1992), "Anticompetitive effects of mergers in markets with localized competition", *Journal of Law, Economics and Organization* 8, 427–440.
- Lommerud, K. and Sorgard, L. (1997), "Mergers and product range rivalry", *International Journal of Industrial Organization* 16, 21–42.
- Lommerud, K., Straume, O. and Sorgard, L. (2005), "Downstream merger with upstream market power", *European Economic Review* 49, 717–743.
- Lommerud, K., Straume, O. and Sorgard, L. (2006), "National versus international mergers in unionized oligopoly", *RAND Journal of Economics* 37 (1), 212–233.
- Long, N.V. and Vouzden, N. (1995), "The effect of trade liberalization on cost reducing horizontal mergers", *Review of International Economics* 3, 141–155.
- Lyons, B.R. (2003), "Could politicians be more right than economists? A theory of merger standards", *European University Institute, Working Paper* 2003/14.
- McAfee, R.P. and Williams, M.A. (1992), "Horizontal mergers and antitrust policy", *Journal of Industrial Economics* 40, 181–187.
- Milliou, C. and Petrakis, E. (2007), "Upstream horizontal mergers, vertical contracts, and bargaining", *International Journal of Industrial Organization* 25, 963–987.
- Milliou, C. and Pavlou, A. (2013), "Upstream mergers, downstream competition and R&D investments", *Journal of Economics and Management Strategy* 22, 787–809.
- Milliou, C. and Sandonís, J. (2014), "Manufacturer mergers and product variety in vertically related markets", *CESifo Working Paper Series* 4932.
- Motta, M. and Vasconcelos, H. (2005). "Efficiency gains and myopic antitrust authority in a dynamic merger game", *International Journal of Industrial Organization* 23, 777–801.
- Neary, J.P. (2007), "Cross-border mergers as instruments of comparative advantage", *Review of Economic Studies* 74, 1229–1257.
- Neven, D. and Röller, L.H. (2005), "Consumer surplus vs. welfare standard in a political economy model of merger control", *International Journal of Industrial Organization* 23 (9–10), 829–848.
- Nilssen, T. and Sorgard, L. (1998), "Sequential horizontal mergers", *European Economic Review* 42, 1683–1702.
- Nocke, V. (2000), "Monopolisation and industry structure", *Economics Working Paper* 2000-W27, Nuffield College, Oxford.
- Nocke, V. and Yeaple, S. (2007), "Cross-border mergers and acquisitions vs. greenfield foreign direct investment: the role of firm heterogeneity", *Journal of International Economics* 72, 336–365.
- Nocke, V. (2008), "Endogenous mergers", in L. Blume and S. Durlauf (eds), *The New Palgrave Dictionary of Economics*, 2nd edition, Basingstoke, UK: Palgrave Macmillan, 578–579.
- Nocke, V. and Whinston, M.D. (2010), "Dynamic merger review", *Journal of Political Economy* 118 (6), 1201–1251.
- Nocke, V. and Whinston, M.D. (2013), "Merger policy with merger choice", *American Economic Review* 103 (2), 1006–1033.
- Nocke, V. and Rey, P. (2014), "Exclusive dealing and vertical integration in interlocking relationships", *CEPR Discussion Paper* DP10176.



- O'Brien, D. and Shaffer, G. (1992), "Vertical control with bilateral contracts", *RAND Journal of Economics* 23 (3), 299–308.
- Perry, M.K. and Porter, R.H. (1985), "Oligopoly and the incentives for horizontal mergers", *American Economic Review* 75, 219–227.
- Pesendorfer, M. (2005), "Mergers under entry", *RAND Journal of Economics* 36 (3), 661–679.
- Rogoff, K. (1985), "The optimal degree of commitment to an intermediate monetary target", *Quarterly Journal of Economics* 100, 1169–1189.
- Salant, S., Switzer, S. and Reynolds, R.J. (1983), "Losses from horizontal merger: the effects of an exogenous change in industry structure on Cournot-Nash equilibrium", *Quarterly Journal of Economics* 98, 185–199.
- Salvo, A. (2010), "Sequential cross-border mergers in models of oligopoly", *Economica* 77, 352–383.
- Sklivas, S. (1987), "The strategic choice of managerial incentives", *RAND Journal of Economics* 18, 452–458.
- Stigler, G. (1950), "Monopoly and oligopoly by merger", *American Economic Review, Papers and Proceedings* 40, 23–34.
- Sutton, J. (1991), *Sunk Costs and Market Structure*, Cambridge, MA: MIT Press.
- Sutton, J. (1998), *Technology and Market Structure: Theory and History*, Cambridge, MA: MIT Press.
- Symeonidis, G. (2010), "Downstream merger and welfare in a bilateral oligopoly", *International Journal of Industrial Organization* 28, 230–243.
- Vasconcelos, H. (2004), "Tacit collusion, cost asymmetries and mergers", *RAND Journal of Economics* 36 (1), 39–62.
- Vasconcelos, H. (2006), "Endogenous mergers in endogenous sunk cost industries", *International Journal of Industrial Organization* 24, 227–250.
- Vives, X. (2002), "Private information, strategic behaviour and efficiency in cournot markets", *RAND Journal of Economics* 33 (3), 361–376.
- Von Ungern-Sternberg, T. (1996), "Countervailing power revisited", *International Journal of Industrial Organization* 14, 507–520.
- Werden, G.J. and Froeb, L.M. (1998) "The entry-inducing effects of horizontal mergers: An explanatory analysis", *Journal of Industrial Economics* 46, 525–543.
- Williamson, O.E. (1968), "Economies as an antitrust defense: The welfare trade-offs", *American Economic Review* 59, 954–959.
- Zhou, V. (2008), "Endogenous horizontal mergers under cost uncertainty", *International Journal of Industrial Organization* 26, 903–912.
- Ziss, S. (1995), "Vertical separation and horizontal mergers", *Journal of Industrial Economics* 43, 63–75.
- Ziss, S. (2001), "Horizontal mergers and delegation", *International Journal of Industrial Organization* 19 (3–4), 471–492.



3. Collusive agreements in vertically differentiated markets

*Marco A. Marini**

1 INTRODUCTION

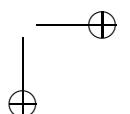
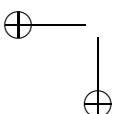
This survey primarily focuses on the incentives of firms to sign collusive agreements in vertically differentiated markets such as, for instance, in cartels, mergers and alliances. It also studies the effects of collusion on market prices and qualities.

The relationship between mergers and price-quality combinations has recently attracted increasing attention in empirical and theoretical industrial organization (IO) literature.¹ On empirical grounds, Berry and Waldfogel (2001) found, for instance, a negative correlation between merging operations and number of existing radio stations with, in addition, an observed increase in radio format varieties related to mergers. Sweeting (2010) and George (2007) reported similar evidence for the US radio music industry and Fan (2013) for the US newspaper market. In airline industries, Peters (2006) observed a reduction of flight frequency in those market segments in which merging carriers compete most, while Mazzeo (2003) showed a deterioration of on-time performance following airline mergers.

In this chapter we introduce a number of game-theoretic tools that can be used to model firm-collusive agreements in vertically differentiated markets. Section 2 quickly reviews the initial literature on price collusion in a vertically differentiated duopoly. Section 3 introduces a vertically differentiated oligopoly setting to study in more detail the incentives of firms to form either the whole market cartel or partial cartels made up of subsets of adjacent firms in the product space, with the aim to collude in prices. This exercise allows us to characterize the price behaviour of alliances by looking, in particular, at the behaviour of what we denote, in turn, *bottom*, *intermediate* and *top* cartels, meaning arbitrary cartels made up of adjacent firms and including either the bottom- or the top-quality firm (in the *bottom* and *top* cartel, respectively) or made up of intermediate firms only (in the *intermediate* cartel). It can be shown that at the price equilibrium for any top or intermediate cartel only two variants remain on sale from the cartel: the highest- and the lowest-quality goods produced by the cartel. On the other hand, in any bottom cartel, only one variant remains on sale, namely the highest quality among those produced *ex ante* by the cartel. The remaining sections focus on the stability of collusion. Section 4, by associating a partition function game with the n -firm vertically differentiated market shows that a sufficient condition for the coalitional stability of the whole industry cartel is the *equidistance* of all firms' qualities. Without this feature, and

* I am very grateful for valuable discussions with and suggestions from Rabah Amir, Maria Rosa Battaggion, Luca Benvenuti, Sergio Currarini, Alberto de Santis, Jean Gabszewicz, Michael Kopel, Sébastien Mitraille, Michele Polo, Giorgio Rodano, Ornella Tarola and the participants at seminars at the University of Milan, Universidad Carlos III de Madrid, Nova Universidade de Lisboa, Université Paris-Dauphine and Università di Roma La Sapienza. I gratefully acknowledge the financial support from CIRIEC International, Liège, Belgium.

¹ Among others, Mazzeo (2002), Crawford and Shum (2006), Gandhi et al., (2008), Draganska, Mazzeo and Seim, (2009), Chu (2010), Byrne (2012), Fan (2013), Lee (2013).



in the presence of highly asymmetric quality gaps, collusive agreements may easily become unstable. Section 5 introduces a standard infinite-horizon game to show that an increase in the number of firms in the market may have contradictory effects on the incentive of firms to collude: collusion may become *easier* for bottom and intermediate firms and *harder* for the top-quality firm. Finally, in Section 6, by means of a three-firm example, I consider the case in which colluding firms can also decide on their quality and price combinations endogenously. In such a case, once merged, firms are allowed to optimally reshape their qualities and prices according to the new market structure. From this, it can be verified whether full or partial cartelizations can be sustained as a subgame perfect equilibria of the whole game, which now includes a coalition formation process taking place at the first stage. For this model we show that partial cartelization always arises in equilibrium with the bottom-quality firm always belonging to the formed cartel. Section 7 concludes.

2 COLLUSION IN A VERTICALLY DIFFERENTIATED DUOPOLY

In his seminal paper, Hackner (1994) analyses the relationship between collusion and vertical product differentiation in an infinitely repeated duopoly framework. The main issue here is to see whether price collusion is more or less likely to be sustained when the quality gap between firms' products is higher. It is shown that the monopoly pricing is more easily attainable when products are closer along the quality ladder. Also, among the two firms, the top-quality firm is the one possessing the highest incentive to break a collusive agreement. This is because with a large quality gap, the profit of the top-quality firm is high even without collusion, and this makes the incentive to collude for this firm weaker than for the bottom-quality firm. In a related paper, Ecchia and Lambertini (1997) study how the stability of price collusion in a vertically differentiated duopoly can be affected by the introduction of a minimum quality standard. The presence of a welfare-maximizing minimum quality standard can make the full collusive agreement harder to sustain. This is because the quality standard decreases the product differentiation, providing the bottom-quality firm with a stronger temptation to defect.²

From the above analyses, two things can be noticed. The first is that, in both models considered above, the degree of product differentiation does not change after a coalition has formed, since the collusive behaviour is restricted to pricing. This assumption is a natural entry point in the literature on cartel stability under product differentiation, as it disentangles the effect of quality gap on the stability of cartels. Further, conceiving collusion in terms of pricing is particularly reasonable from a short-run perspective. Still, it leaves a companion question unexplored, namely the effect of the cartel on product differentiation. This analysis could be particularly pertinent in a long-run perspective since one cannot exclude the fact that in a more extended time span, a coalition (typically a cartel or a merger) entails structural changes, such as relocation of production facilities, or adjustment in the product range and quality.

The second is, instead, that in both papers the market is duopolistic and, as a result, any cooperation between the two firms implies by definition a full market cartelization. There are remarkable examples where firms form partial alliances (i.e. those including a subset

² The contradictory results among the two papers mainly depend on their different cost assumptions.

of firms in the market) rather than a whole market coalition. Actually, in partial alliances, colluding firms can still compete against rival firms outside the coalition, and the effects of partial alliances or mergers are not equivalent to those observed when all firms mimic the behaviour of a monopolist.

Lambertini (2000) explores how cartel stability can be connected to the R&D activity in a duopoly in which the collusive quality choice may occur either under price- or quantity-setting behaviour.³ The issue concerning alliance formation with more than two firms in a vertically differentiated market remains, however, unexplored, as does the effect of partial collusion on market equilibrium. Scarpa (1998) models a vertical differentiation market with three firms competing in quality and prices.⁴ In particular, he considers the role of a minimum quality standard, and highlights how the demand level of each firm in a vertically differentiated market only depends on quality and price of adjacent firms in the product space. Indeed, since only adjacent variants compete against each other, under partial collusion defining the optimal set of products to market requires balancing the *cannibalization effect* that a variant produced by the coalition exerts *within* the coalition with the possibility that this variant *steals* consumers from the rival firms (*stealing effect*).

Other related papers are those by Lommerud and Sorgard (1997), Gandhi et al. (2008), Chen and Schwartz (2013) and Brekke, Siciliani and Straume (2014), all devoted to the analysis of price-quality post-merger repositioning.⁵ Lommerud and Sorgard (1997) is inspired by Salant, Switzer and Reynolds (1983) and Deneckere and Davidson (1985) and it is devoted to evaluating the profitability of a merger under both Cournot and Bertrand competition. The authors assume that the market is initially populated by three firms and, therefore, two firms can merge and decide on the number of brands to market. When the fixed cost of marketing a brand is “high”, the merged entity reduces its product range. This increases the profitability of mergers both under Bertrand and Cournot competition due to reduced marketing costs. With a “low” cost of marketing, the effect on the product range depends both on the nature of competition and on the degree of product differentiation. For example, under Cournot or Bertrand competition and sufficiently differentiated products, the non-merging firm finds it profitable to introduce a new brand, thereby damaging the merged entity. In order to highlight the impact of a merger on non-price competition, Gandhi et al. (2008) assume instead that firms can instantaneously and costlessly reposition their products after a merger, thereby choosing both price and location in a Hotelling market. They show that after a merger the products are repositioned away from each other to reduce the resulting cannibalization effect. Consequently, non-merging substitutes are repositioned between the merged products and, after all these location strategies, the merged firm’s incentive to raise prices decreases. Similarly, in a Hotelling framework, Chen and Schwartz (2013) analyse the incentive for firms to introduce a product innovation when proposing a merger-to-monopoly. In contrast with Arrow’s (1962) finding for process innovation, where the monopolist never undertakes R&D efforts to innovate, in this paper the incentive to invest in incremental

³ A different strand of literature considers the possible impacts of R&D joint ventures on product market collusion. See on this, Martin (1995), Lambertini, Poddar and Sasakic (2002) and Marini, Petit and Sestini (2014).

⁴ Recently, Pezzino (2010) has developed the same model under quantity competition. Cesi (2010) studies the effect of two-firm mergers in a three-firm market in the presence of a social welfare-maximizing minimum quality standard.

⁵ Other recent papers by Mazzeo (2002), Einav (2003) and Seim (2006) look at the price-quality strategies decided on by industry entrants.

product innovations can be higher for the merged entity (a monopolist) than for a rival facing competition from the existing good. Indeed, the monopolist can coordinate the pricing of the two products overcompensating for the erosion of profits coming from cannibalization. In a spatial competition model à la Salop with three *ex ante* identical firms, Brekke et al. (2014) show that any two-firm merger reduces its product quality, whereas the non-merging firm responds by increasing its quality. Final prices can either increase or decrease according to the responsiveness of demand functions. Moreover, it is shown that if a merger entails the closure of one of the two merged firms, this always leads to higher qualities and prices for all firms in the market.

3 COLLUSION IN AN N -FIRM VERTICALLY DIFFERENTIATED MARKET

As underlined above, although easily interpretable, a two-firm vertically differentiated market possesses a few limitations and does not allow a full-fledged analysis of market partial cartelization. Therefore, in this first modelling section we simply extend a traditional model à la Mussa and Rosen (1978) and Gabszewicz and Thisse (1979) to an n -firm market in order to see the main implications in terms of pricing behaviour under collusion.⁶

Let n firms $k = 1, 2, \dots, n$ supply n different quality variants q_1, q_2, \dots, q_n with $q_k \in [0, \infty]$ and $q_n > q_{n-1} > \dots > q_1$ to a population of consumers. As in Mussa and Rosen (1978) and Gabszewicz and Thisse (1979) consumers are indexed by θ and uniformly distributed in the interval $[0, \beta]$, with $\beta < \infty$. As usual, the parameter θ captures consumers' willingness to pay for quality: the higher θ , the higher the baseline utility gained when consuming variant q_k of the product. Each consumer can either buy one unit of a variant or not buy at all. Formally, a simple way to represent consumer's utility is

$$U(\theta) = \begin{cases} \theta q_k - p_k \text{ when buying variant } k \\ 0 \text{ when not buying.} \end{cases} \quad (3.1)$$

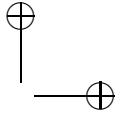
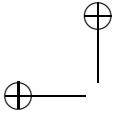
where p_k is the price set by firm k , such that $p_k \in [0, \bar{p}]$, where $0 < \bar{p} < \infty$ is a given upper bound on prices. From the above formulation, the marginal consumer buying variant $k = 1$ is

$$\theta_1 = \frac{p_1}{q_1},$$

and the market is partially *uncovered*, with some consumers excluded from buying even the bottom-quality variant. In general, the consumer indifferent between buying variant $k - 1$ and k for $k = 2, 3, \dots, n$ is

$$\theta_k = \frac{p_k - p_{k-1}}{q_k - q_{k-1}}.$$

⁶ In their seminal paper Gabszewicz and Thisse (1980) introduce an n -firm model of vertically differentiated firms under the assumption of *equispaced* products.



38 Handbook of game theory and industrial organization: applications

where $p_k > p_{k-1}$ for every $k = 1, 2, 3, \dots, n$. For the time being, we assume that product qualities are exogenously given and we disregard costs to simplify calculations.⁷

When considering price competition, the payoffs of all firms can be easily characterized by describing the payoff of three types of firms in the quality spectrum: (i) top quality, (ii) intermediate quality and (iii) bottom quality. The top-quality firm (denoted $k = n$) sets a price p_n to maximize its profit

$$\Pi_n = D_n p_n = \left(\beta - \frac{p_n - p_{n-1}}{q_n - q_{n-1}} \right) p_n. \quad (3.2)$$

Conversely, every intermediate firm $k = 2, 3, \dots, n-1$ maximizes

$$\Pi_k = D_k p_k = \left(\frac{p_{k+1} - p_k}{q_{k+1} - q_k} - \frac{p_k - p_{k-1}}{q_k - q_{k-1}} \right) p_k. \quad (3.3)$$

Finally, the bottom-quality firm ($k = 1$), maximizes

$$\Pi_1 = D_1 p_1 = \left(\frac{p_2 - p_1}{q_2 - q_1} - \frac{p_1}{q_1} \right) p_1. \quad (3.4)$$

The optimal reply of every non-cooperative firm can be easily obtained as follows:

$$p_n(p_{n-1}) = \frac{1}{2} (p_{n-1} + \beta(q_n - q_{n-1})) \quad (3.5)$$

for the *top-quality* firm ($k = n$),

$$p_k(p_{k-1}, p_{k+1}) = \frac{1}{2} \frac{p_{k-1}(q_{k+1} - q_k) + p_{k+1}(q_k - q_{k-1})}{(q_{k+1} - q_{k-1})} \quad (3.6)$$

for every *intermediate-quality* firm $k = 2, 3, \dots, n-1$, and

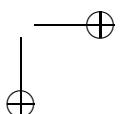
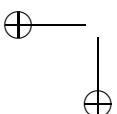
$$p_1(p_2) = \frac{1}{2} \frac{p_2 q_1}{q_2} \quad (3.7)$$

for the *bottom-quality* firm ($k = 1$).

Expressions (3.2)–(3.4) show that prices and qualities are strategic complements for all firms ($\frac{\partial^2 \Pi_k}{\partial p_k \partial q_k} > 0$) and the best reply of every firm shifts outward due to an increase in its quality. On the other hand, for every firm k , an increase in the quality of direct rivals' products q_j , for $j = (k+1)$ and $(k-1)$ causes a negative effect on its profit ($\frac{\partial \Pi_k}{\partial p_k \partial q_j} < 0$) and price competition becomes tougher as a result. Note also that, from (3.2)–(3.4), all firms' profit functions are concave in their own prices and also their choice sets are compact and convex and their best replies are *contractions*,⁸ in such a way that the existence of a

⁷ The existence of quality fixed costs does not alter the nature of the results obtained here.

⁸ See Gabszewicz, Marini and Tarola (2017).



unique (non-cooperative) Nash equilibrium n -price vector p^* associated with the n variants (q_1, q_2, \dots, q_n) is guaranteed for any (finite) number of firms competing in the market.⁹

3.1 Full Price Collusion

When firms form the whole market cartel, they can be assumed to maximize the sum of all firms' payoffs:

$$\Pi_{\{N\}} = \sum_{k=1}^n \Pi_k = \Pi_1 + \dots + \Pi_{k-1} + \Pi_k + \Pi_{k+1} + \dots + \Pi_n.$$

For every colluding firm $k = 1, \dots, n$, the first-order condition is written as¹⁰

$$\frac{\partial \Pi_{\{N\}}}{\partial p_k} = \frac{\partial \Pi_{k-1}}{\partial p_k} + \frac{\partial \Pi_k}{\partial p_k} + \frac{\partial \Pi_{k+1}}{\partial p_k} = 0. \quad (3.8)$$

Since the top-quality firm $k = n$ in the cartel internalizes the payoff of its lower-quality neighbour, its optimal reply is written as

$$p_n^c(p_{n-1}) = p_{n-1} + \frac{\beta}{2} (q_n - q_{n-1}). \quad (3.9)$$

From the same rationale, for all intermediate firms $k = 2, 3, \dots, (n-1)$ that are members of the cartel, the optimal reply is written as

$$p_k^c(p_{k-1}, p_{k+1}) = \frac{p_{k-1}(q_{k+1} - q_k) + p_{k+1}(q_k - q_{k-1})}{(q_{k+1} - q_{k-1})}, \quad (3.10)$$

since they internalize the payoff of their adjacent neighbour members of the cartel. Finally, the optimal reply of the bottom-quality firm $k = 1$ is given by

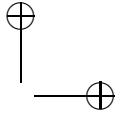
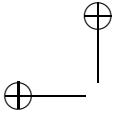
$$p_1^c(p_2) = \frac{q_1}{q_2} p_2. \quad (3.11)$$

As already pointed out by Gabszewicz et al. (1986) and, more recently, by Gabszewicz, Marini and Tarola (2017), in a model in which unit costs vary only mildly with quality, under full price collusion the n firms set prices p_k^c such that their market shares are nil for all firms except for the top-quality one ($k = n$). In particular, under full collusion, for every firm $k = 1, 2, \dots, n$ and $j < k$, it is easy to obtain prices as

$$p_k^c = \frac{1}{2} \beta \sum_{j=1}^k (q_j - q_{j-1}). \quad (3.12)$$

⁹ See, for instance Friedman (1991), p. 84.

¹⁰ Note that $\frac{\partial^2 \Pi_N}{\partial p_i^2} = -\frac{2(v_{i+1} - v_{i-1})}{(v_{i+1} - v_i)(v_i - v_{i-1})} < 0$ for $i = 2, 3, \dots, n-1$, and, therefore, the joint profit Π_N is concave in every firm's price p_i . The same condition holds for the two extreme firms along the quality spectrum, i.e. $i = 1$ and $i = n$.



Inserting (3.12) into every firm's market share D_k , we obtain for the bottom-quality firm,

$$D_1(p_1^c, p_2^c) = \left(\frac{p_2^c - p_1^c}{\tau_2} - \frac{p_1^c}{\tau_1} \right) = \left(\frac{\frac{1}{2}\beta(\tau_1 + \tau_2) - \frac{1}{2}\beta\tau_1}{\tau_2} - \frac{\frac{1}{2}\beta\tau_1}{\tau_1} \right) = 0$$

where $\tau_j = (q_j - q_{j-1})$ denotes the quality gap of every firm j selling goods of lower or equal quality than firm k , and $\tau_1 = (q_1 - q_0) = q_1$. Moreover, inserting (3.12) into every *intermediate-quality* firm's market share τ_k , we obtain:

$$\begin{aligned} D_k(p_{k-1}^c, p_k^c, p_{k+1}^c) &= \left(\frac{p_{k+1}^c - p_k^c}{\tau_{k+1}} - \frac{p_k^c - p_{k-1}^c}{\tau_{k-1}} \right) = \\ &= \left(\frac{\frac{1}{2}\beta \sum_{j \leq k+1} \tau_j - \frac{1}{2}\beta \sum_{j \leq k} \tau_j}{\tau_{k+1}} - \frac{\frac{1}{2}\beta \sum_{j \leq k} \tau_j - \frac{1}{2}\beta \sum_{j \leq k-1} \tau_j}{\tau_{k+1}} \right) = \\ &= \left(\frac{\frac{1}{2}\beta \tau_{k+1}}{\tau_{k+1}} - \frac{\frac{1}{2}\beta \tau_k}{\tau_k} \right) = 0. \end{aligned}$$

with,

$$\begin{aligned} D_n(p_{n-1}^c, p_n^c) &= \left(\beta - \frac{p_n^c - p_{n-1}^c}{q_n - q_{n-1}} \right) = \left(\beta - \frac{\frac{1}{2}\beta \sum_{j \leq n} \tau_j - \frac{1}{2}\beta \sum_{j \leq n-1} \tau_j}{\tau_n} \right) = \\ &= \left(\beta - \frac{\frac{1}{2}\beta \tau_n}{\tau_n} \right) = \frac{1}{2}\beta, \end{aligned}$$

for the *top-quality* firm. Thus, when colluding together all firms cover only half of the market and the whole market payoff is:

$$\Pi_{\{N\}} = \sum_{k \in N} \Pi_k^{(N)} = \sum_{k=1}^n p_k^c D_k = \frac{1}{4} \beta^2 \tau_n. \quad (3.13)$$

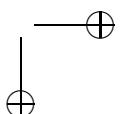
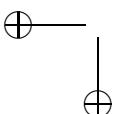
3.2 Partial Cartels

In many cases firms can organize themselves in a coalition structure (partition) of the N firms different from the grand coalition, $C = (S_1, S_2, \dots, S_m)$, with $m \leq n$. However, in a vertically differentiated market every firm can effectively distort prices by colluding either with its left (lower quality) or right (higher quality) or with both its local competitors.¹¹ In what follows we introduce a few simple definitions to develop the analysis of partial cartelization. In order to affect prices, firms can form *bottom-*, *intermediate-* or *top-quality* cartels. For each of these members, the first-order condition of profit maximization is written as follows:

- (i) In the case of *interior* cartel members:

$$\frac{\partial \Pi_S}{\partial p_k} = \frac{\partial \sum_{k \in S} \Pi_k}{\partial p_k} = \frac{\partial \Pi_{k-1}}{\partial p_k} + \frac{\partial \Pi_k}{\partial p_k} + \frac{\partial \Pi_{k+1}}{\partial p_k} = 0,$$

¹¹ Price collusion can also occur among disconnected firms, but in this case the prices of the firms will just be equal to those arising at the non-cooperative equilibrium.



leading to the optimal reply function

$$p_k^{pc}(p_{k-1}, p_{k+1}) = \frac{p_{k-1}(q_{k+1} - q_k) + p_{k+1}(q_k - q_{k-1})}{(q_{k+1} - q_{k-1})}, \quad (3.14)$$

where the superscript *pc* stands for *partial collusion*.

- (ii) In the case of *lower boundary* cartel member:

$$\frac{\partial \Pi_S}{\partial p_k} = \frac{\partial \sum_{k \in S} \Pi_k}{\partial p_k} = \frac{\partial \Pi_k}{\partial p_k} + \frac{\partial \Pi_{k+1}}{\partial p_k} = 0,$$

leading to the best-reply function

$$p_k^{pc}(p_{k-1}, p_{k+1}) = \frac{\frac{1}{2}p_{k-1}(q_{k+1} - q_k) + p_{k+1}(q_k - q_{k-1})}{(q_{k+1} - q_{k-1})}. \quad (3.15)$$

- (iii) Finally, in the case of *upper boundary* cartel member:

$$\frac{\partial \Pi_S}{\partial p_k} = \frac{\partial \sum_{k \in S} \Pi_k}{\partial p_k} = \frac{\partial \Pi_k}{\partial p_k} + \frac{\partial \Pi_{k-1}}{\partial p_k} = 0,$$

leading to the best-reply function

$$p_k^{pc}(p_{k-1}, p_{k+1}) = \frac{p_{k-1}(q_{k+1} - q_k) + \frac{1}{2}p_{k+1}(q_k - q_{k-1})}{(q_{k+1} - q_{k-1})}. \quad (3.16)$$

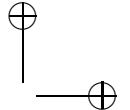
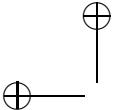
Definition 1 (i) A bottom cartel $S_B \subset N$ is a coalition formed by consecutive intermediate firms $k = 2, \dots, n-1$, also including the bottom-quality firm $k = 1$. (ii) An intermediate cartel $S_k \subset N$ is a coalition formed by at least two consecutive intermediate firms $k = 2, \dots, n-1$. (iii) A top cartel $S_T \subset N$ is a coalition formed by consecutive intermediate firms $k = 2, \dots, n-1$, also including the top-quality firm $k = n$.

Following Gabszewicz et al. (2017), the next proposition characterizes the market shares of firms belonging to (i) an intermediate cartel; (ii) a bottom cartel; (iii) a top cartel:

Proposition 1 (i) A bottom cartel only produces in equilibrium the top-quality variant among those in the cartel. (ii) Any intermediate cartel only produces in equilibrium the top- and the bottom-quality variants among those in the cartel. (iii) Any top cartel only produces in equilibrium the top- and the bottom-quality variants among those in the cartel.

Proof See the Appendix. ■

Corollary In a generic partition of the n firms $P = (S_1, S_2, \dots, S_m)$ organized in $m < n$ non-trivial cartels, a total of $2m + (n - z) - 1$ (resp. $2m + (n - z)$) variants are put on sale in the market when the partition includes (resp. does not include) the bottom cartel, for $z = s_1 + s_2 + \dots + s_m$, where s_j , for $j = 1, 2, \dots, m$, denotes the cardinality of every cartel.



42 Handbook of game theory and industrial organization: applications

In order to complete the characterization of every partial cartelization of the market we can provide a price comparison for all firms under partial cartelization with respect to both fully non-cooperative and fully collusive cases:

Proposition 2 *Under partial cartelization the firms set prices p_k^{pc} higher or equal to the corresponding prices p_k^* charged at the non-cooperative price equilibrium and lower than the corresponding full collusive prices p_k^c .*

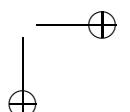
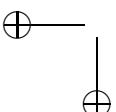
Proof Let us assume, for simplicity, that only one cartel $S \subset N$ has formed, and that the remaining firms play as singletons. Note, however, that the same reasoning would apply to the case with more than one cartel. It can easily be verified that the joint profit of an arbitrary cartel Π_S is continuous and concave with respect to every firm's price p_k , for $k \in S$. Moreover, the optimal reply of every partially collusive firm $k \in S$ is a *contraction* and, hence, a unique partially collusive price profile p^{pc} exists for any given level of qualities q_1, q_2, \dots, q_n , under the assumption that all firms with unsold goods set their equilibrium prices exactly at the levels for which the sales become nil.¹² Thus, we can: (a) start with a profile p^* of Nash equilibrium prices, (b) let the firms in $S \subset N$ reply according to their optimal collusive replies. A quick comparison between optimal replies under partial cartelization (3.14)–(3.16) and purely non-cooperative Nash equilibrium shows that the former are always steeper than the latter and, since they are in both cases positively sloped, all firms in the cartel will set higher prices than in the non-cooperative scenario. (c) Similarly, an increase in prices will also occur in all firms in the fringe playing non-cooperatively: given the higher prices of the cartel, they will respond, in turn increasing their prices. (d) The described adjustment process, given the contraction property of all firms' optimal replies, will converge to a new profile of prices such that $p_k^{pc} > p_k^*$ for all $k = 1, 2, \dots, n$. Inequality $p_k^c > p_k^{pc}$, for every $k = 1, 2, \dots, n$, can be proved along similar lines. ■

4 A COOPERATIVE APPROACH TO THE STABILITY OF THE WHOLE INDUSTRY AGREEMENT

In this section we consider the incentive of firms to form a whole industry cartel (grand coalition). Following Gabszewicz et al. (2016), a partition function game can be associated with the vertically differentiated market introduced in Section 3 and, from this, it can be proved that the core of this game is non-empty when the qualities of the products sold by the firms are equispaced along the quality spectrum. Moreover, it can be easily shown that, when this regularity condition does not hold, the core can be empty. Therefore, a fully collusive agreement among firms is more easily reachable when there are neither too large nor too asymmetric gaps between firms' qualities. The symmetry in quality gaps helps to maintain the discipline of the whole market cartel because it reduces the incentive of firms to free-ride by leaving the agreement.

Following Gabszewicz et al. (2016) we adopt the concept of *delta core* by Hart and Kurz (1983), also denoted *projection core* in the recent axiomatization by Bloch and Van den Nouweland (2014). Since for the case of vertically differentiated markets the coalitional worth

¹² Any higher price would be equally optimal since these goods are not purchased by consumers.



possesses positive coalition externalities,¹³ the delta or projection core is the smallest core and, therefore, its existence implies the existence of all other possible versions of core in games with simultaneous moves.¹⁴

To prove our result we can formally associate with the vertically differentiated market described above a *partition function game* $P = (N, v(S, C))$, where N is the set of firms and $v(S, C) \in \mathcal{R}_+$ is the worth associated with every coalition of firms $S \subset N$ belonging to a coalition structure $C \in \mathcal{C}$. In our model, when a cartel $S \subset N$ forms, its maximal coalitional payoff is obtained when the remaining firms in $N \setminus S$ stick together in the complementary coalition $\{N \setminus S\}$. Therefore, if the core of the partition function game P exists when the coalitional worth $v(S, C)$ is computed for $C = (S, N \setminus S)$, it will *a fortiori* exist for any other partition of the firms in $N \setminus S$.

Definition 2 *The core of partition function game $P = (N, v(S, C))$ consists of all efficient payoff allocations $\Pi \in \mathcal{R}_+^{|N|}$ respecting $\sum_{k \in S} \Pi_k \geq v(S, C)$ for all $S \subset N$ and for all partition $C \in \mathcal{C}$ to which S may belong.*

Then we have the following result:

Proposition 3 *Let market variants q_1, q_2, \dots, q_n be equispaced with $(q_k - q_{k-1}) = \tau \in [0, \infty]$ for all $k = 1, 2, \dots, n$, with $q_0 = 0$. Then, the core of the partition function game $P = (N, v(S, C))$ associated with the n -firm vertically differentiated market is non-empty.*

The proof of this result, contained in Gabszewicz et al. (2016) is constructive and it finds a specific allocation of the monopoly profit $\Pi^{\{N\}}$ such that neither an individual firm nor a bottom, intermediate or top cartel have an incentive to leave the grand coalition under its maximal expectation, i.e. that the remaining firms continue to collude inside the complementary cartel $N \setminus S$. Such allocation is simply

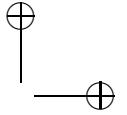
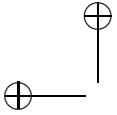
$$\Pi = (s_1 \Pi^{\{N\}}, s_2 \Pi^{\{N\}}, \dots, s_n \Pi^{\{N\}}), \quad (3.17)$$

where s_k for $k = 1, 2, \dots, n$ are shares of the monopoly profit given by

$$s_k = \frac{\Pi^{\{k, N \setminus k\}}}{\sum_{k \in N} \Pi^{\{k, N \setminus k\}}},$$

¹³ This means that every firm is advantaged when rivals merge in coalitions.

¹⁴ Gabszewicz et al. (2016) use this notion of core in order to provide the strongest core existence result. Demange (1994) provides general conditions for core existence in economies producing differentiated goods, although in absence of externalities between coalitions. Zhao (2013) examined the existence of α -, γ - and δ -core in a three-firm linear Cournot oligopoly with different marginal costs. In a differentiated quantity oligopoly with three (or four firms) Watanabe and Matsubayashi (2013) show that for any degree of product differentiation the γ -core is non-empty while the δ -core only exists in the presence of high product differentiation. For a more detailed account of the work dealing with coalitional agreements in oligopoly games see Marini (2009) and Curarini and Marini (2015).



44 Handbook of game theory and industrial organization: applications

such that $\sum_{k \in N} s_k = 1$, where $\Pi_{\{k\}}^{\{k, N \setminus k\}}$ is the profit of every firm k when in competition with its complementary coalition $N \setminus \{k\}$.

As the simple example below shows, when the quality gaps among firms widely differ, the core can be empty.

4.1 An Empty Core Example

Let us assume four firms in the market, i.e. $N = \{1, 2, 3, 4\}$, initially selling four different qualities q_1, q_2, q_3, q_4 . Let now the firms fully collude by forming the grand coalition. Let now the top cartel $S_T = \{2, 3, 4\}$ decide to leave the grand coalition and coalition structure $C = (\{1\}, \{2, 3, 4\})$ forms as a result. In this case, the top cartel obtains:

$$\Pi_T^{(\{1\}, \{2, 3, 4\})} = \frac{\beta^2 q_2 q_3 (q_3 - q_2)}{(4q_3 - q_2)^2} + \frac{1}{4} \frac{\beta^2 (4q_2 q_4 - q_1 q_4 - 3q_1 q_2)}{(4q_2 - q_1)}.$$

For $\beta = 1$, $q_1 = 1$, $q_2 = 5$ and $q_4 = 10$ and $q_3 > 7.26$, the quality gap between q_2 and q_3 (both produced inside the cartel) becomes sufficiently high for

$$\Pi_T^{(\{1\}, \{2, 3, 4\})} + \Pi_1^{(\{1\}, \{2, 3, 4\})} > \Pi_N^{\{N\}} = \frac{1}{4} \beta^2 q_4 = 2.5,$$

and the core is, therefore, empty. If all products are equispaced, with $q_1 = 2.5$, $q_2 = 5$, $q_3 = 7.5$ and $q_4 = 10$,

$$\Pi_1^{(\{1\}, \{2, 3, 4\})} + \Pi_T^{(\{1\}, \{2, 3, 4\})} = 2.21 < \Pi_N^{\{N\}}.$$

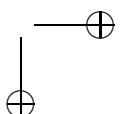
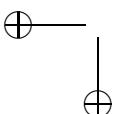
Moreover, it can be verified that all other feasible deviations by singletons or coalitions of firms do not improve upon the grand coalition allocations. Core existence is, in such a way, guaranteed.

5 A NON-COOPERATIVE APPROACH TO THE STABILITY OF THE WHOLE INDUSTRY AGREEMENT

In this section we test the stability of the whole industry cartel using a standard repeated-game approach. For this purpose, we use the model with equispaced variants, which is sufficiently tractable.¹⁵

We already obtained in Section 3 the monopoly payoff. What is required to characterize the standard *grim* strategy of a standard infinite-horizon extension of the vertically differentiated model is to make explicit the non-cooperative equilibrium payoffs of all firms and, as a second aspect, to define their *defection* payoffs obtained when playing their best replies when all other rivals collude. Finally, an intuitive *allocation rule* has to be introduced to divide the fully collusive payoff among the n heterogeneous firms. In what follows we derive the price vector

¹⁵ In this section we use part of the material contained in Bos and Marini (2016).



obtained at the Nash equilibrium of every *constituent* game under the equispaced product assumption:

Proposition 4 Let market variants q_1, q_2, \dots, q_n be equispaced and such that $q_k - q_{k-1} = \tau$ for every $k = 1, 2, \dots, n$, with $q_0 = 0$. Then, the non-cooperative Nash equilibrium price vector for all firms $k = 1, 2, \dots, n$ is given by:

$$p_k^* = \frac{\tau\beta(b_1^k - b_2^k)}{\sqrt{3}b_1^n + \sqrt{3}b_2^n},$$

and for $b_1 = (2 + \sqrt{3})$ and $b_2 = (2 - \sqrt{3})$.

Proof See the Appendix. ■

If we assume the existence of quadratic quality costs for each firm $c(q_k) = \frac{q_k^2}{2}$, their non-cooperative payoffs can be written as

$$\Pi_k^* = \left(\frac{p_{k+1}^* - p_k^*}{\tau} - \frac{p_k^* - p_{k-1}^*}{\tau} \right) p_k^* - \frac{q_k^2}{2} = \frac{2}{3} \frac{\tau\beta^2(b_1^k - b_2^k)^2}{(b_1^n + b_2^n)^2} - \frac{(\tau k)^2}{2}.$$

Now, since the fully collusive price under equally spaced variants is, for every firm $k = 1, 2, \dots, n$

$$p_k^c = \frac{1}{2}\beta\tau k,$$

we can easily characterize the fully collusive profit of every firm (before any transfer takes place) as

$$\Pi_k^c = \frac{\beta^2\tau(k)}{4} - \frac{(\tau k)^2}{2}.$$

One way to divide the fully collusive profit among all firms is to use the following natural quality ranking:

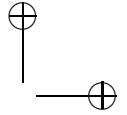
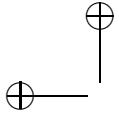
$$r_k = k \cdot \tau$$

for every $k = 1, 2, \dots, n$, which substantially corresponds to the position of each firm in the equispaced quality space. Therefore, using the fact that

$$\sum_{k=1,\dots,n} r_k = \frac{n(n+1)\tau}{2}$$

at the fully collusive agreement we can simply assign to every firm a personalized share equal to:

$$\alpha_k = \frac{r_k}{\sum_{k=1,\dots,n} r_k} = \frac{2k}{n(n+1)}.$$



46 *Handbook of game theory and industrial organization: applications*

Finally, using every firm's non-cooperative best replies we can easily obtain every firm's defection profit as:

$$\begin{aligned}\Pi_1^d &= \left(\frac{p_2^c - 2p_1^d}{\tau} \right) p_k^d - \frac{\tau^2}{2} = \left(\frac{\frac{1}{2}\beta\tau(2)-2\left(\frac{1}{4}\tau\beta\right)}{\tau} \right) \frac{1}{4}\tau\beta - \frac{\tau^2}{2} = \frac{1}{8}\tau\beta^2 - \frac{\tau^2}{2}, \\ \dots \\ \Pi_k^d &= \left(\frac{p_{k+1}^c - 2p_k^d + p_{k-1}^c}{\tau} \right) p_k^d - \frac{(\tau(k))^2}{2} = \frac{(k)^2\tau\beta^2}{8} - \frac{(\tau(k))^2}{2}, \\ \dots \\ \Pi_n^d &= \left(\beta - \frac{p_n^d - p_{n-1}^c}{\tau} \right) p_n^d - \frac{(\tau n)^2}{2} = \frac{(n+1)^2\tau\beta^2}{16} - \frac{(\tau n)^2}{2}.\end{aligned}$$

Thus, for the full collusion to be sustained as a subgame perfect Nash equilibrium of the infinite-horizon game (via a grim strategy) the discount factor of every firm has to respect the following condition:

$$\delta_k(\beta, \tau, n) \geq \frac{\Pi_k^d - \alpha_k \Pi_k^c}{\Pi_k^d - \Pi_k^*} = \frac{\left(\frac{(k)^2}{8} \tau \beta^2 - \frac{(\delta(k))^2}{2} \right) - \frac{2(k)}{n(n+1)} \left(\frac{\beta^2 \tau n}{4} - \frac{(\tau n)^2}{2} \right)}{\frac{(k)^2}{8} \beta^2 \tau - \frac{2}{3} \frac{\beta^2 \tau ((b_1)^k - (b_2)^k)^2}{((b_1)^n + (b_2)^n)^2}} \quad (3.18)$$

for all $k = 1, 2, \dots, n - 1$ and

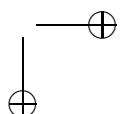
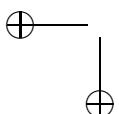
$$\delta_n(\beta, \tau, n) \geq \frac{\Pi_n^d - \alpha_n \Pi_n^c}{\Pi_n^d - \Pi_n^*} = \frac{\left(\frac{n^2 \beta^2 \tau}{8} - \frac{(\tau n)^2}{2} \right) - \frac{2n}{n(n+1)} \left(\frac{\beta^2 \tau n}{4} - \frac{(\tau n)^2}{2} \right)}{\frac{(n+1)^2}{16} \tau \beta^2 - \frac{2}{3} \frac{\tau \beta^2 ((b_1)^n - (b_2)^n)^2}{((b_1)^n + (b_2)^n)^2}} \quad (3.19)$$

¹⁶ From the above expressions, the following proposition results:

Proposition 5 Let market variants q_1, q_2, \dots, q_n be equispaced with $q_k - q_{k-1} = \tau$ for every $k = 1, 2, \dots, n$ and $q_0 = 0$. Let also every firm's share of the monopoly profit be determined by its quality ranking, as $\alpha_k = 2(k)/n(n+1)$. Then, an increase in the number of firms n reduces the discount factor sustaining the fully cooperative agreement as a subgame perfect Nash equilibrium of the infinitely repeated game via a grim strategy for all firms $k = 1, 2, 3, \dots, n-1$, while it increases the discount factor of the top-quality firm $k = n$ (for $n > 3$).

Proof This can be obtained from straightforward manipulations of expressions (3.10)–(3.11). ■

¹⁶ Note that a constraint for $\beta > \sqrt{n\delta}\sqrt{2}$ must be imposed for both collusive and non-cooperative firms' payoffs to be non-negative.



Proposition 5 helps us to see that, in vertically differentiated markets, under equispaced variants, an increase in the number of firms has contradictory effects on the incentive of firms to collude: it makes collusion *easier* to sustain for bottom- and intermediate-quality firms but, at the same time, it makes it harder for the top-quality firm. This result is somehow surprising if compared to the usual view that collusion is more easily sustainable when the number of firms is small, whereas it usually becomes harder when the number of firms increases.

6 MERGERS AND ALLIANCES WITH ENDOGENOUS QUALITIES

To the best of our knowledge, a full-fledged theoretical study of the effects of alliances and mergers on market prices and qualities in a *vertically* differentiated industry with more than two firms has not yet been provided. Similarly unexplored is the analysis of *merger stability* between firms in vertically differentiated markets when firms can reshape prices and qualities of the products after mergers. Anecdotal evidence shows that mergers and acquisitions often occur among firms selling fairly differentiated products along the quality spectrum. For instance, some of the mergers taking place after the 1979 deregulation of the US airline market, occurred between one big national/international carrier and one low-fare local carrier (e.g. the merger between American Airlines and AirCal in 1986 or between Delta and Atlantic Southeast Airlines in 1999)¹⁷ or, alternatively, just between intermediate-quality carriers (as for Southwest Airlines and AirTran Airways in 2010 or between Republic Airways and Midwest Airlines in 2009). Analogously, the European airline industry has witnessed a high number of mergers occurring between broadly differentiated airlines such as, for instance, between Air France and Air Inter in 1999 or between Lufthansa and Air Dolomiti in 2003.¹⁸

In a similar vein, the automotive industry has plenty of examples of premium car producers taking over economy car manufacturers, as in the merger between Volkswagen Group and Skoda in 1991 or between Nissan and Renault in 1999.

One consequence of these consolidation processes is often whether to reposition the lower-quality brand towards a higher segment of the market or, in some other cases, to *unbrand* intermediate-quality products to create a *fighting brand* able to compete more aggressively with the firms positioned at the bottom of the quality spectrum. However, the latter strategy usually appears more of a temporary than a permanent strategy, since a fighting brand may risk cannibalizing the market of the merging firms. Ultimately, a consolidated group can find it more advantageous to *rebrand* its economy products rather than *unbrand* some of its intermediate-quality outlets. Instead of letting Smart ForTwo cars compete in the low segment of the market, Daimler-Benz preferred to transform this city car into a premium car. Similarly, the boom of mergers recently observed in pharmaceutical industries, involving top pharmaceutical companies acquiring generic drugs manufacturers (as in the recent case of Teva absorbing Allergan Generics), could represent a similar trend.¹⁹

In Gabszewicz, Marini and Tarola (2015) we introduce a simple framework in which three *ex ante* heterogeneous firms, initially producing three vertically differentiated goods, *low*

¹⁷ See: <http://www.airlines.org>.

¹⁸ In some other cases the low-cost airlines have attempted to take over small-medium airlines, as in the recent hostile takeover bid launched by Ryanair for Aer Lingus.

¹⁹ See, for instance, Wieczner (2015).

(firm 1), *medium* (firm 2) and *high* (firm 3), enter a negotiation to decide whether to merge or not with some or all rival firms and, once merged, optimally reshape the qualities and prices according to the new market structure.

Assume as in Gabszewicz et al. (2015) a three-stage game where, at the first stage, every firm expresses its willingness to form an alliance or, alternatively, to stay as a singleton. Then, at the second and third stage each formed coalition can decide, in turn, the qualities and prices of its goods. An alliance can either contain all firms in the market (*grand coalition*), as $N = \{1, 2, 3\}$ or, alternatively, be formed by a non-empty subset $S \subset N$ of firms, with $S \in \mathcal{N}$, where $\mathcal{N} = 2^N \setminus \{\emptyset\}$ is the set of all non-empty coalitions of the N firms:

$$\mathcal{N} = (\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}).$$

Thus, the set \mathcal{C} of all coalition structures C that can be formed by the three firms is:

$$\mathcal{C} = ((\{1\}, \{2\}, \{3\}), (\{1, 2\}, \{3\}), (\{1\}, \{2, 3\}), (\{1, 3\}, \{2\}), (\{1, 2, 3\})).$$

The game can be solved backwards to analyse the prices and qualities selected in equilibrium by firms under the assumption that either the grand coalition or any other intermediate coalition structure has formed at the first stage. As in Bloch (1995, 1996) and Ray and Vohra (1999), the coalition formation game can be assumed *sequential*, with an exogenous order of play. Different from these authors, since firms are *ex ante* heterogeneous, it is assumed that every firm proposes not only an alliance, but also a division of the coalition payoff. Each recipient of the proposal can either accept or reject the offer and, in the case of rejection, it becomes its turn to make a proposal. The game is assumed finite horizon and every firm only possesses one turn to make a proposal at each period.²⁰

Since prices and qualities are selected in a sequence by every formed coalition, the payoffs accruing to a firm or a coalition in each feasible coalition structure can be easily obtained as shown in Table 3.1:

Table 3.1 Firm payoffs in every coalition structure

Coalition structure	Payoffs		
(\{1\}, \{2\}, \{3\})	$\Pi_1^* = 0.00005\beta^4$	$\Pi_2^* = 0.00124\beta^4$	$\Pi_3^* = 0.02348\beta^4$
(\{1, 2, 3\})	$\Pi_{(\{1,2,3\})}^{(N)} = 0.03125\beta^4$		
(\{1, 2\}, \{3\})	$\Pi_{\{1,2\}}^{(\{1,2\},\{3\})} = 0.00152\beta^4$	$\Pi_3^{(\{1,2\},\{3\})} = 0.02443\beta^4$	
(\{1, 3\}, \{2\})	$\Pi_{\{1,3\}}^{(\{1,3\},\{2\})} = 0.02443\beta^4$	$\Pi_2^{(\{1,3\},\{2\})} = 0.00152\beta^4$	
(\{1\}, \{2, 3\})	$\Pi_1^{(\{1\},\{2,3\})} = 0.00152\beta^4$	$\Pi_{\{2,3\}}^{(\{1\},\{2,3\})} = 0.02443\beta^4$	

²⁰ Both Bloch's (1995, 1996) and Ray and Vohra's (1999) models are an infinite-horizon negotiation process.

It turns out that, although the qualities and prices arising in each partial merger do not vary, the profits accruing to firms depend on the coalitions to (against) which they belong (compete).

Moreover, using the above payoffs, it can be shown that, although the full monopolization of the market is the most profitable outcome of the game, in a finite-horizon sequential game of coalition formation the incentive for firms to enter a whole industry merger is dominated by that to form partial mergers. In particular, the finite-horizon sequential coalition formation game reaches the results described by the following proposition:

Proposition 6 (i) When the high-quality firm 3 is the initiator of the sequential coalition formation game, the only stable coalition structure is $C_{12,3} = (\{1,2\}, \{3\})$, where firm 3 continues to produce the top variant q_3 and the two remaining firms 1 and 2 only market intermediate variant q_2 . (ii) When firm 2 is the initiator of the game, the only stable coalition structure is $C_{13,2} = (\{1,3\}, \{2\})$, where firm 1 and 3 jointly produce top variant q_3 and firm 2 produces intermediate variant q_2 . (iii) Finally, when firm 1 is the initiator of the game, the only stable coalition structure is $C_{12,3} = (\{1,2\}, \{3\})$, where firm 3 produces top variant q_3 and 1 and 2 jointly produce intermediate variant q_2 .

Proof See Gabszewicz et al. (2015). ■

Notice that, both in (i) and (ii) the initiator of the game never belongs to an alliance in equilibrium. Indeed, the payoff of a firm when it remains singleton (rationally expecting that the other firms will prefer to merge) dominates that of being part of the grand coalition, since in the latter case the distribution of profits would be unfavourable to the initial proposer. The equilibrium profit accruing to either firm 2 or 3 when initiating the game and competing against an alliance is, therefore, larger than when they are part of the alliance. The optimal strategy is, therefore, to induce the remaining firms to merge. A different result arises when firm 1 (the bottom-quality one) begins the negotiation process. In this case, firm 1 cannot credibly commit to remain independent since the remaining firms (2 and 3) prefer to play as singletons rather than forming an alliance (see Table 3.1). This is due to the fact that the alliance between firm 2 and 3 is problematic since in this circumstance 2 would optimally leapfrog the bottom-quality firm, ending up by sharing the top-quality firm's duopoly payoff, which is lower than the sum of the firms' profits under triopoly. Knowing in advance the infeasibility of coalition $\{2,3\}$, firm 1 would prefer to let firm 3 play independently and, then, form an alliance with firm 2.

A striking result of this model is that all equilibrium mergers always contain the bottom-quality firm, which, in all cases, drops its low-quality variant from the market. In particular, whoever is the additional player in a coalition (either the intermediate- or the top-quality firm), equilibrium prices and qualities always coincide with that observed in the case of a duopoly, with a high-quality firm competing against a low-quality rival, as in Motta (1992).

At first sight, this result seems to be counterintuitive. A natural conjecture would be that either the range of variants or the quality gap between variants in the market would change with the players involved in the alliance. This model shows instead that only profits accruing to the single firms change with the type of partial merger, range of products, quality gap and price being unchanged. Indeed, the cannibalization effect and the stealing effect induce the merger, whatever its members, to withdraw from the market the lowest-quality variant between the set that can be produced *a priori*. Interestingly, depending on the intensity of

these effects, in some circumstances this variant is withdrawn from the market at the price stage; in other circumstances at the quality stage. In particular, the merger formed by the intermediate-quality and by the low-quality firm stops immediately to market the bottom-quality product at the price stage. In contrast, the merger formed by the top- and the bottom-quality firm keeps the bottom product (as a fighting brand) at the price stage, whereas it ultimately drops it at the quality stage. As argued above, keeping a fighting brand in an alliance is mostly a short-run (price) than a medium-/long-run (quality) strategy and it is, therefore, dropped when the merging group can reposition its product lines. Finally, it is found that, in all equilibrium (partial) mergers, the bottom-quality firm is always present. This appears to be in line with numerous theoretical and experimental studies on coalition formation in triads of heterogeneous individuals, i.e. possessing different skills or fighting ability (e.g. Caplow, 1956, 1959, 1968, Vinacke and Arkoff, 1957, Gamson, 1961). A central conclusion of these studies is that “weakness is strength” (see, for instance, Mesterton-Gibbons et al., 2011, p. 189), meaning that less-powered individuals usually have more opportunity to be part of a coalition.

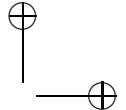
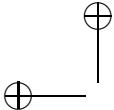
The results of this coalition formation game confirms that the most likely mergers occur between intermediate- and bottom-quality producers, with the premium-quality brands preferably running alone. This is the case with some top car producers (such as, for instance, Daimler-Benz) whose only participation is in a few specific projects. What the model results also indicate is that mergers between intermediate- and bottom-quality firms, as occurred between Volkswagen and Skoda, or between Fiat and Chrysler in the automotive industry, should be the norm. In these cases the intermediate-quality product is withdrawn from the market, which can be interpreted as meaning that all products sold by the merger have a tendency to converge towards the same level of quality of their premium brand products. The model also stresses how mergers between top- and bottom-quality firms are likely, such as, for instance, those that recently occurred between generic pharmaceutical manufacturers and premium brand pharmaceutical companies. The model results suggest that in this case the bottom-quality products can be profitably retired from the market to soften the competition among remaining goods.

7 CONCLUDING REMARKS

The rationale underlying many of the results presented in this chapter can be found in the nature of competition among vertically differentiated firms. Indeed, in any cartel or merger, the optimal set of products to market is defined by balancing the *cannibalization effect* within the coalition with the *stealing effect* occurring between a coalition and the firms outside. It was shown that the bottom-quality variant in a group of colluding firms is kept on sale in the market only when such a cartel needs it as a sort of *fighting brand* to protect itself from all lower-quality variants sold by the fringe of competitors. In any other case a cartel prefers to withdraw all its low-quality variants from the market. In this way firms can soften price competition in the market and magnify the quality differentiation between the variants remaining on sale. This view seems in line with the empirical findings, where mergers emphasize “product differentiation” among merging firms as well as with respect to their outside rivals. Partial mergers can therefore be viewed as a means to enhance the *dynamic* competition for the market and to reduce the *static* competition in the market.

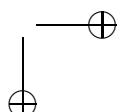
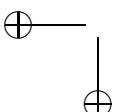
REFERENCES

- Arrow, K. (1962), "Economic Welfare and the Allocation of Resources for Invention", in NBER (ed.), *The Rate and Direction of Inventive Activity*, Princeton, NJ: Princeton University Press, pp. 609–626.
- Berry, S. and J. Waldfogel (2001), "Do Mergers Increase Product Varieties?" *Quarterly Journal of Economics*, 116, 1009–1025.
- Bloch, F. (1995), "Endogenous Structures of Associations in Oligopolies", *Rand Journal of Economics*, 26, 537–556.
- Bloch, F. (1996), "Sequential Formation of Coalitions with Fixed Payoff Division", *Games and Economic Behavior*, 14, 90–123.
- Bloch, F. and A. Van den Nouweland (2014), "Expectation Formation Rules and the Core of Partition Function Games", *Games and Economic Behavior*, 88, 339–353.
- Bos, I. and M.A. Marini (2016), "Sustaining Price Collusion in Vertically Differentiated Oligopolies", mimeo.
- Brekke, K.R., L. Siciliani and O.R. Straume (2014), "Horizontal Mergers and Product Quality", *NHH Discussion Paper*, February.
- Byrne, D.P. (2015), "The Impact of Consolidation on Cable TV Prices and Product Quality", *International Economic Review*, 56, 805–850.
- Caplow, T. (1956), "A Theory of Coalitions in the Triad", *American Sociological Review*, 21, 489–493.
- Caplow, T. (1959), "A Theory of Coalitions in the Triad", *American Journal of Sociology*, 64, 488–493.
- Caplow, T. (1968), *Two Against One: Coalitions in Triads*, Englewood Cliffs, NJ: Prentice-Hall.
- Cesi, B. (2010), "Mergers Under Minimum Quality Standard: A Note", *Economics Bulletin*, 30, 4, 3260–3266.
- Chen, Y. and M. Schwartz (2013), "Product Innovation Incentives: Monopoly vs. Competition", *Journal of Economics & Management Strategy*, 22, 3, 513–528.
- Chu, C.S. (2010), "The Effect of Satellite Entry on Cable Television Prices and Product Quality", *RAND Journal of Economics*, 41, 730–764.
- Crawford, G.S. and M. Shum (2006), "The Welfare Effects of Endogenous Quality Choice: The Case of Cable Television", mimeo, University of Arizona.
- Curarini, S. and M.A. Marini (2015), "Coalitional Approaches to Collusive Agreements in Oligopoly Games", *The Manchester School*, 83, 3, 253–287.
- Demange, G. (1994), "Intermediate Preferences and Stable Coalition Structures", *Journal of Mathematical Economics*, 23, 45–58.
- Deneckere, R. and C. Davidson (1985), "Incentive to Form Coalitions with Bertrand Competition", *Rand Journal of Economics*, 16, 473–486.
- Draganska, M., M. Mazzeo and K. Seim (2009), "Beyond Plain Vanilla: Modeling Joint Product Assortment and Pricing Decisions", *Quantitative Marketing and Economics*, 7, 105–146.
- Ecchia, G. and L. Lambertini (1997), "Minimum Quality Standards and Collusion", *Journal of Industrial Economics*, 45, 1, 101–113.
- Einav, L. (2010), "Not all Rivals Look Alike: Estimating an Equilibrium Model of the Release Date Timing Game", *Economic Inquiry*, 48, 2, 369–390.
- Fan, Y. (2013), "Ownership Consolidation and Product Characteristics: A Study of the U.S. Daily Newspaper Market", *American Economic Review*, 103, 1598–1628.
- Friedman, J.W. (1989), *Game Theory with Applications to Economics*, Oxford: Oxford University Press.
- Gabszewicz, J.J. and J.-F. Thisse (1979), "Price Competition, Quality and Income Disparities", *Journal of Economic Theory*, 20, 340–359.
- Gabszewicz, J.J. and J.-F. Thisse (1980), "Entry (and Exit) in a Differentiated Industry", *Journal of Economic Theory*, 22, 327–338.
- Gabszewicz, J.J., M.A. Marini and O. Tarola (2015), "Alliance Formation in Vertically Differentiated Markets", *CORE Discussion Papers*, 30/2015, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
- Gabszewicz, J.J., M.A. Marini and O. Tarola (2016), "Core Existence in Vertically Differentiated Markets", *Economics Letters*, 149, 28–32.
- Gabszewicz, J.J., M.A. Marini and O. Tarola (2017), "Vertical Differentiation and Collusion: Proliferation or Cannibalization?", *Research in Economics*, 71, 1, 129–139.
- Gabszewicz, J.J., A. Shaked, J. Sutton and J.-F. Thisse (1986), "Segmenting the Market: The Monopolist's Optimal Product Mix", *Journal of Economic Theory*, 39, 2, 273–289.
- Gamson, W.A. (1961), "A Theory of Coalitions in the Triad", *American Sociological Review*, 21, 489–493.
- Gandhi, A., L. Froeb, S. Tschanz and G. Werden (2008), "Post-merger Product Repositioning", *Journal of Industrial Economics*, 56, 49–67.
- George, L. (2007). "What's Fit to Print: The Effect of Ownership Concentration on Product Variety in Daily Newspapers Markets", *Information Economics and Policy*, 19, 285–303.
- Hackner, J. (1994), "Collusive Pricing in Markets for Vertically Differentiated Products", *International Journal of Industrial Organization*, 12, 2, 155–177.



52 *Handbook of game theory and industrial organization: applications*

- Hart, S. and D. Myatt (1983), "Endogenous Formation of Coalitions", *Econometrica*, 52, 1047–1064.
- Lambertini, L. (2000), "Technology and Cartel Stability Under Vertical Differentiation", *German Economic Review*, 1, 4, 421–444.
- Lambertini, L., S. Poddar and D. Sasakic (2002), "Research Joint Ventures, Product Differentiation, and Price Collusion", *International Journal of Industrial Organization*, 20, 829–854.
- Lee, J. (2013), "Endogenous Product Characteristics in Merger Simulation: A Study of the U.S. Airline Industry", mimeo.
- Lommerud, K. and L. Sorgard (1997), "Merger and Product Range Rivalry", *International Journal of Industrial Organization*, 16, 1, 21–42.
- Marini, M.A. (2009), "Games of Coalition & Network Formation: A Survey", in A.K. Naimzada, S. Stefani and A. Torriero (eds), *Networks, Topology and Dynamics. Lectures Notes in Economics & Mathematical Systems*, 613, 67–93, Berlin: Springer-Verlag.
- Marini, M.A., M.L. Petit and R. Sestini (2014), "Strategic Timing in R&D Agreements", *Economics of Innovation and New Technology*, 23, 3, 274–303.
- Martin, S. (1995), "R&D Joint Ventures and Tacit Product Market Collusion", *European Journal of Political Economy*, 11, 733–741.
- Mazzeo, M. (2002), "Product Choice and Oligopoly Market Structure", *Rand Journal of Economics*, 33, 221–242.
- Mazzeo, M. (2003), "Competition and Service Quality in the U.S. Airline Industry", *Review of Industrial Organization*, 22, 275–296.
- Mesterton-Gibbons, M., S. Gavrilov, G. Janko and E. Akcay (2011), "Models of Coalition or Alliance Formation", *Journal of Theoretical Biology*, 274, 187–204.
- Motta, M. (1992), "Cooperative R&D and Vertical Product Differentiation", *International Journal of Industrial Organization*, 10, 643–661.
- Mussa, M. and S. Rosen (1978), "Monopoly and Product Quality", *Journal of Economic Theory*, 18, 301–317.
- Peters, C. (2006), "Evaluating the Performance of Merger Simulation: Evidence from the U.S. Airline Industry", *Journal of Law and Economics*, 49 (2), 627–649.
- Pezzino, M. (2010), "Minimum Quality Standards with More Than Two Firms Under Cournot Competition", *The IUP Journal of Managerial Economics*, 8, 3, 26–45.
- Ray, D. and R. Vohra (1999), "A Theory of Endogenous Coalition Structures", *Games and Economic Behavior*, 26, 2, 286–336.
- Salant, S.W., S. Switzer and R.J. Reynolds (1983), "Losses from Horizontal Merger: The Effects of an Exogenous Change in Industry Structure on Cournot-Nash Equilibrium", *Quarterly Journal of Economics*, 98, 2, 185–199.
- Scarpa, C. (1998), "Minimum Quality Standards with More Than Two Firms", *International Journal of Industrial Organization*, 16, 5, 665–676.
- Seim, K. (2006), "An Empirical Model of Firm Entry with Endogenous Product-type Choices", *Rand Journal of Economics*, 37, 619–640.
- Sweeting, A. (2010), "The Effects of Mergers on Product Positioning: Evidence from the Music Radio Industry", *Rand Journal of Economics*, 41, 372–397.
- Vinacke, V.E. and A. Arkoff (1957), "An Experimental Study of Coalitions in the Triad", *American Sociological Review*, 22, 406–414.
- Watanabe, T. and N. Matsubayashi (2013), "Note on Stable Mergers in Markets with Asymmetric Substitutability", *Economics Bulletin*, 33, 2024–2033.
- Wieczner, J. (2015), "The Real Reasons for the Pharma Merger Boom", *Fortune*, July 28.
- Zhao, J. (2013), "The Most Reasonable Solution for an Asymmetric Three-firm Oligopoly", mimeo, Saskatchewan, Canada, March.



APPENDIX: OMITTED PROOFS

Proof of Proposition 1 (Gabszewicz et al., 2017). Take a generic *intermediate* cartel of $h \leq n - 2$ firms initially selling variants

$$q_k, q_{k+1}, q_{k+2}, \dots, q_{k+h}$$

and competing, both with a left-hand fringe of independent firms selling lower-quality variants q_1, q_2, \dots, q_{k-1} and with a right-hand fringe selling, alternatively, higher-quality variants $q_{k+h+1}, q_{k+h+2}, \dots, q_n$. Using expressions (3.14)–(3.16) the optimal replies of the firms in the cartel are

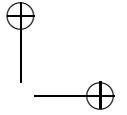
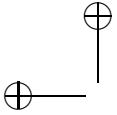
$$\begin{aligned} p_k^{pc}(p_{k-1}, p_{k+1}) &= \frac{\frac{1}{2}p_{k-1}(q_{k+1} - q_k) + p_{k+1}(q_k - q_{k-1})}{(q_{k+1} - q_{k-1})} \\ p_{k+1}^{pc}(p_k, p_{k+2}) &= \frac{p_k(q_{k+2} - q_{k+1}) + p_{k+2}(q_{k+1} - q_k)}{(q_{k+2} - q_k)} \\ p_{k+2}^{pc}(p_{k+1}, p_{k+3}) &= \frac{p_k(q_{k+3} - q_{k+2}) + p_{k+3}(q_{k+2} - q_{k+1})}{(q_{k+3} - q_{k+1})} \\ &\dots \\ p_{k+h}^{pc}(p_{k+h-1}, p_{k+h+1}) &= \frac{p_{k+h-1}(q_{k+h+1} - q_{k+h}) + \frac{1}{2}p_{k+h+1}(q_{k+h} - q_{k+h-1})}{q_{k+h+1} - q_{k+h-1}}. \end{aligned}$$

where only the two extreme firms k and $k + h$ in the cartel are directly competing with the firms outside. Without loss of generality, take a generic firm inside the cartel producing an intermediate variant (i.e. neither the bottom- nor the top quality within the cartel), say firm $k + 1$. Using both the optimal reply of firm $k + 1$ and those of the firms connected to it (i.e. firms k and $k + 2$) and rearranging, we obtain the optimal replies of these three firms as functions of p_{k-1} and p_{k+3} only:

$$\begin{aligned} \tilde{p}_k &= p_k^{pc}(p_{k-1}, p_{k+3}) = \frac{1}{2} \frac{p_{k-1}(q_{k+3} - q_k) + 2p_{k+3}(q_k - q_{k-1})}{q_{k+3} - q_{k-1}}, \\ \tilde{p}_{k+1} &= p_{k+1}^{pc}(p_{k-1}, p_{k+3}) = \frac{1}{2} \frac{p_{k-1}(q_{k+3} - q_{k+1}) + 2p_{k+3}(q_{k+1} - q_{k-1})}{q_{k+3} - q_{k-1}}, \\ \tilde{p}_{k+2} &= p_{k+2}^{pc}(p_{k-1}, p_{k+3}) = \frac{1}{2} \frac{p_{k-1}(q_{k+3} - q_{k+2}) + 2p_{k+3}(q_{k+2} - q_{k-1})}{q_{k+3} - q_{k-1}}. \end{aligned}$$

Using the above, we can easily compute the optimal market share of firm $(k + 1)$ as

$$D_{k+1}(\tilde{p}_k, \tilde{p}_{k+1}, \tilde{p}_{k+2}) = \frac{\tilde{p}_{k+2} - \tilde{p}_{k+1}}{q_{k+2} - q_{k+1}} - \frac{\tilde{p}_{k+1} - \tilde{p}_k}{q_{k+1} - q_k} = 0,$$



54 Handbook of game theory and industrial organization: applications

which proves that under partial collusion every intermediate firm of an *intermediate* cartel obtains zero market share. Repeating now the same procedure for the firm producing the lowest quality in the cartel (here firm k), we obtain instead that

$$D_k(\tilde{p}_k, \tilde{p}_{k+1}, \tilde{p}_{k-1}) = \frac{\tilde{p}_{k+1} - \tilde{p}_k}{q_{k+1} - q_k} - \frac{\tilde{p}_k - \tilde{p}_{k-1}}{q_k - q_{k-1}} = \frac{1}{2} \frac{\tilde{p}_{k-1}}{(q_k - q_{k-1})} > 0$$

for $\tilde{p}_{k-1} > 0$. Finally, computing the optimal replies of the highest-quality firm in the cartel, i.e. firm $(k+h)$, and of the firms directly connected to it, we obtain

$$\begin{aligned} \tilde{p}_{k+h-1}(p_{k+h-2}, p_{k+h}) &= \frac{p_{k+h-2}(q_{k+h-1} - q_{k+h-2}) + p_{k+h}(q_{k+h-1} - q_{k+h-2})}{q_{k+h} - q_{k+h-2}} \\ \tilde{p}_{k+h}(p_{k+h-1}, p_{k+h+1}) &= \frac{p_{k+h-1}(q_{k+h+1} - q_{k+h}) + \frac{1}{2}p_{k+h+1}(q_{k+h} - q_{k+h-1})}{q_{k+h+1} - q_{k+h-1}} \\ \tilde{p}_{k+h+1}(p_{k+h}, p_{k+h+2}) &= \frac{1}{2} \frac{p_{k+h}(q_{k+h+2} - q_{k+h+1}) + p_{k+h+2}(q_{k+h+1} - q_{k+h})}{q_{k+h+2} - q_{k+h}}, \end{aligned}$$

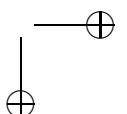
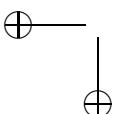
Using the above,

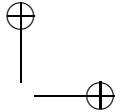
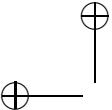
$$\begin{aligned} D_{k+h}(\tilde{p}_{k+h-1}, \tilde{p}_{k+h}, \tilde{p}_{k+h+1}) &= \frac{\tilde{p}_{k+h+1} - \tilde{p}_{k+h}}{q_{k+h+1} - q_{k+h}} - \frac{\tilde{p}_{k+h} - \tilde{p}_{k+h-1}}{q_{k+h} - q_{k+h-1}} = \\ &= \frac{1}{2} \frac{\tilde{p}_{k+h+1}}{(q_{k+h} - q_{k+h-1})} > 0, \end{aligned}$$

showing that only the variants produced by the two firms at the extremes of this (generic) intermediate cartel are sold at prices implying *positive* market shares. Exactly the same procedure proves that, in a *top cartel*, only the highest- and the lowest-quality variants initially sold by the cartel remain on sale.

Finally, let us consider a *bottom cartel*, i.e. a cartel formed by firms $1, 2, \dots, h$ initially selling h variants q_1, q_2, \dots, q_h and competing with $(n-h)$ independent firms selling the higher-quality variants $q_{h+1}, q_{h+2}, \dots, q_n$. Again, we can apply the same argument used above to show that every firm in the *interior* of the cartel (i.e. neither selling the lowest-quality nor the highest-quality variant in the cartel) obtains zero market share. Also, for the top-quality firm in the cartel (here firm h), we obtain that $D_h(\tilde{p}_h, \tilde{p}_{h-1}, \tilde{p}_{h+1}) > 0$. Finally, when considering a firm selling the lowest-quality variant in any *bottom* cartel, its market share is simply written as:

$$D_1(p_2, p_1) = \frac{p_2 - p_1}{q_2 - q_1} - \frac{p_1}{q_1},$$





which, using firm 1's optimal collusive reply $p_1^{pc}(p_2) = \frac{q_1}{q_2}p_2$, becomes

$$D_1(p_2, \tilde{p}_1) = \frac{p_2 - \frac{q_1}{q_2}p_2}{q_2 - q_1} - \frac{\frac{q_1}{q_2}p_2}{q_1} = 0,$$

showing that, differently from other cartels, a *bottom cartel* optimally produces only its top-quality variant q_h . **Q.E.D.**

Proof of Proposition 4 Under equispaced variants, from (3.6), for all $k = 1, 2, \dots, n$ best replies are

$$p_k = \frac{1}{4}(p_{k+1} + p_{k-1}),$$

which can be written as a second-order difference equation as

$$p_{k+1} - 4p_k + p_{k-1} = 0,$$

with complementary function

$$Ab^{k+1} - 4Ab^k + Ab^{k-1} = 0$$

and whose associated characteristic function possesses two distinct real roots given by

$$b_1 = 2 + \sqrt{3}, b_2 = 2 - \sqrt{3},$$

implying

$$p_k = A_1 b_1^k + A_2 b_2^k. \quad (\text{A1})$$

Moreover, using the fact that for the bottom-quality firm,

$$p_1 = \frac{1}{4}p_2 = \frac{1}{4}(p_2 + p_0)$$

we can set

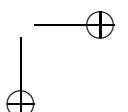
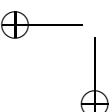
$$p_0 = A_1 b_1^0 + A_2 b_2^0 = A_1 + A_2 = 0,$$

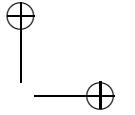
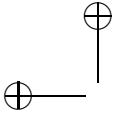
implying

$$A_2 = -A_1. \quad (\text{A2})$$

Finally, using the fact that for the top-quality firm

$$p_n = \frac{1}{2}(p_{n-1} + \beta\tau)$$





56 *Handbook of game theory and industrial organization: applications*

we just write

$$2p_n - p_{n-1} = \beta\tau,$$

which implies

$$p_{n-1} = A_1 b_1^{n-1} + A_2 b_2^{n-1} = A_1 (b_1^{n-1} - b_2^{n-1}) + 2A_1 (b_1^n - b_2^n) - \beta\tau$$

from which

$$A_1 (b_1^{n-1} - b_2^{n-1}) - 2A_1 (b_1^n - b_2^n) + \beta\tau = 0$$

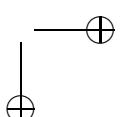
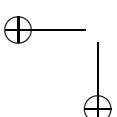
and, then,

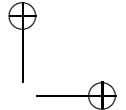
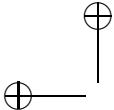
$$A_1 = \frac{\beta\tau}{(2b_1^n - 2b_2^n - b_1^{n-1} + b_2^{n-1})}.$$

As a final step, we insert coefficients A_1 and A_2 into (A1), obtaining

$$p_k^* = A_1 (b_1)^k + A_2 (b_2)^k = A_1 (b_1)^k - A_1 (b_2)^k = \frac{\beta\tau (b_1^k - b_2^k)}{\sqrt{3}b_1^n + \sqrt{3}b_2^n},$$

for every $k = 1, 2, \dots, n$ and $b_1 = (2 + \sqrt{3})$ and $b_2 = (2 - \sqrt{3})$, which concludes the proof. **Q.E.D.**





4. Cartels and leniency: Taking stock of what we learnt

Giancarlo Spagnolo and Catarina Marvão

1 INTRODUCTION

Basic law and economics explains that, as with other forms of public law enforcement, anti-cartel enforcement increases social welfare if the social gain from detecting and deterring cartels is larger than the deadweight loss from wasteful administration, prosecution and litigation activities (e.g., Posner, 1976; Cooter and Ulen; 1988), and from the possible economic distortions directly caused by poorly designed enforcement policies (like fines based on firms' turnover; see Bageri, Katsoulacos and Spagnolo, 2013).

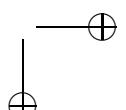
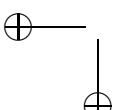
Law enforcement agencies, and more specifically competition authorities, have started to publish in their annual reports the number of successful cartel convictions and the amount of fines collected, implicitly proposing them as performance measures. While this increase in transparency is welcomed, it should also be taken with due care, as it may tend to generate a discrepancy between the objectives of law enforcement agencies and those of society. Using the number of cases, of successful convictions, or of fines collected as a measure of output or performance creates a natural incentive to win many easy cases, possibly abusing leniency policy (and plea bargaining) by being too generous, so as to win more cases more easily. An overly generous leniency policy offering fine reductions to several reporting firms may make a competition authority appear very successful in terms of the number of cases won, of firms convicted, or amount of fines collected, while reducing social welfare by decreasing cartel deterrence (because firms expect a lenient treatment if caught) and increasing the amount of prosecution costs (because there are more prosecuted cartels).¹

Moreover, when considering firms' incentives to apply for leniency, we are unavoidably drawn into discussing the issue of cartel damages, and more generally of the interaction between private and public law enforcement. The recently approved EU Damages Directive is likely to significantly change the incentives for applying for leniency in Europe, and, potentially, in directions that differ depending on the previously existing national legislation.

For welfare considerations, there may be a need to further increase (e.g., criminal) sanctions and step up proactive cartel detection policies (e.g., incentivizing whistleblowers and screening large databases) not as possible substitutes but as potential complements to less generous leniency policies. This combination of tools may improve efficiency and social welfare by increasing cartel deterrence and reducing the large deadweight loss society currently suffers in terms of private and public litigation costs.

In the past five years, 29 international cartels have been discovered in the EU and the USA, including automotive parts' suppliers, which are the largest set of bid-rigging schemes ever

¹ If all firms could obtain almost full amnesty by self-reporting under a leniency policy, then all cartel members would all self-report all the time, the competition authority would generate and win a lot of cases, but cartel formation would go up, together with enforcement expenditures.



discovered, suggesting that antitrust enforcement still has limited deterrence effects (of course, we are not arguing that full cartel deterrence would be optimal, just that deterrence seems to be limited).

There is little evidence on the effects of fines on corporate governance, such as CEO disqualification and clauses in the contracts of CEOs, lower management or others. This is particularly important in the banking sector where financial stability concerns generate a “too-big-to-fine” problem. A recent example is from the euro interest rate derivatives (EIRD) cartel.² The CEO of Barclays Bank, Bob Diamond, was fired for applying for leniency, while the CEOs of the other banks remained in place (or received large severance packages). This illustrates the large gap between the current climate, where colluding managers are rewarded (revealed preferences argument), and a scenario with optimal enforcement. The fact that there is little evidence of colluding managers being punished and managerial contracts having provisions against collusion, suggests that European Commission (EC) fines are too low, given that even in the absence of EC criminal sanctions, they are not harsher than US fines.

It can also be observed that some of these cartels are convicted in an increasing number of jurisdictions in parallel. There are clear signs that the perceived “pros” of leniency are leading the EC to a particularly generous use of the leniency tool, almost as if it was a form of plea bargaining: leniency reductions have been granted to 52 percent of all EC cartel fines (1998–2014), and this percentage, corresponding to an average of four leniency recipients per cartel, is on the rise. Extended leniency seems to be driven by an attempt to solve a problem – too many cartel cases waiting to be prosecuted by competition authorities with a fixed amount of resources – which may be worsened, rather than solved, by overusing leniency to speed up cases but thereby reducing cartel deterrence. The negative impact of a possible excessive use of leniency, on the other hand, is substantial for society, which must bear the deadweight loss from the large administration and prosecution costs of all the cases, in addition to those of cartels that are not deterred.

The widespread adoption and use of leniency policies brought with it a vast and rapidly growing strand of economic research. Theoretical studies have sought to examine the effects of these programs, as tools to enhance the detection, prosecution and deterrence of cartel conduct. The research has highlighted the strong potential for well-designed and well-managed leniency policies to contribute to social welfare, but it has also highlighted the serious risk of poorly implemented leniency programs, which may intensify cartel formation and increase enforcement costs at the same time.

Of course, once theory clarifies the trade-offs, finding “the right amount of leniency” becomes a mainly empirical issue, so that availability of the necessary data may be a serious problem. Indeed, the secrecy of collusive agreements poses several challenges for economic research. Recent studies have tried to empirically test the effectiveness of leniency policies, particularly those administered by the US Department of Justice (DOJ) and the EC, given the current level of sanctions. This research has considerable potential value in assisting competition authorities to design optimal policies by having a better understanding of the impact that such policies, their specific features and manner of administration, have on the behavior of cartel participants.

² The cartel took place between September 2005 and May 2008 and it involved seven banks (Barclays, Crédit Agricole, HSBC, JPMorgan Chase, Deutsche Bank, RBS and Société Générale) over varying time periods and covering the whole European Economic Area.

However, indirect methods based on strong assumptions need to be employed to infer whether an increase or decrease in the number of convicted cartels observed after a policy change is due to better enforcement, or due to an increase in the number of cartels present in society. Therefore, evaluating the deterrence effects of leniency policies (and of the level of sanctions) remains very difficult, as cartels are not readily observable in society unless they are convicted. This makes laboratory experiments an important complementary tool that some researchers have used to test different hypothetical policies that have not actually been implemented. The drawback of these studies is that they are always subject to stronger external validity caveats than empirical studies, particularly when used to approximate firm behavior. However, in the case of cartels and analogous crimes, laboratory experiments are particularly valuable and recent work on collusive corruption by Armantier and Boly (2013) seems to suggest that external validity concerns may not be too troublesome.

The rest of this chapter develops as follows. In Section 2, we review the theoretical studies of leniency, with a particular focus on post-2008 studies. In Section 3 we examine empirical evidence on leniency policies, having regard to both descriptive and econometric studies. In Section 4, we review the experimental evidence. Section 5 concludes.

2 THEORETICAL STUDIES ON LENIENCY

2.1 Early Theoretical Studies

Let us briefly recall a few main results from the “early” literature, before 2008 (we can only consider a few of these studies here; a detailed review of this early literature is available in Spagnolo, 2008). The law enforcement literature analyzed leniency and self-reporting, focusing on individual wrongdoers committing occasional crimes (Malik, 1993; Kaplow and Shavell, 1994; Innes, 1999). In particular, Kofman and Lawarrée (1996) offer the first model of how collusion in a hierarchy can be prevented by leniency. However, these papers present static models, mostly of single agent crimes, which cannot capture deterrence originated by the dynamic effects of leniency on cartels. The plea-bargaining literature is also related to leniency, since it explores the efficiency of fine reductions in exchange for post-prosecution cooperation from cartel members (Grossman and Katz, 1983; Reinganum, 1988; Kobayashi, 1992). However, this literature cannot capture the *ex ante* effect of leniency on wrongdoers who have not been prosecuted.

The theoretical literature on leniency programs is vast and growing due to the increasing use of leniency in the EU and the USA and its adoption in other jurisdictions. The pioneering paper by Motta and Polo (2003) was the first to address the effect of leniency on cartels, in a dynamic analytical framework, focused on the effects of leniency programs on the prosecution stage and the allocation of the budget of a competition authority. The crucial result of the model is that in a second-best scenario, in which the competition authority needs to introduce leniency policies because of lack of internal resources, the positive effect of leniency on deterrence, through faster, cheaper (freeing up resources for cartel detection) and more effective prosecution, tends to dominate the negative effect on deterrence, that is, the reduction of overall sanctions. This first model assumed, to simplify, that if a cartel member defects, it immediately becomes not liable anymore, and therefore cannot account for the incentives

firms may have to report when they want to leave/deviate from a cartel, when in reality a deviating cartel member remains liable for several years; nor for the incentives to report first, generated by leniency and linked to the fear of being sanctioned because somebody else reports instead. By disregarding these forces (crucial according to experimental evidence), this first model led to a series of other results/implications that clashed with the intuition of practitioners and legal scholars, including that leniency before an investigation cannot have deterrence effects – it only has effects if it is available after an investigation has been opened; that all cartel members should be awarded leniency if they apply, not only first applicants; and that leniency is second-best, and should not be introduced if the competition authority has a sufficiently large budget.

By assuming away the ability of well-administered leniency policies to induce non-detected cartels to come forward and self-report before an investigation is opened, which is instead the core aspect that distinguishes leniency policies from plea bargaining (see, e.g., Lewis, 2006), this first model focused entirely on leniency for already detected cartels and the prosecution phase. Also, its policy prescriptions to award leniency to many applicants after the cartels' discovery were rather divergent from the DOJ's view on the crucial features of the Corporate Leniency Policy, 1993 (Hammond, 2004b). To improve on all these grounds, Spagnolo (2000a, 2004), Rey (2003) and Aubert, Kovacic and Rey (2006) developed models that, on the contrary, focused on the potential direct incentives' effects of leniency highlighted by legal scholars and the DOJ but not accounted for in Motta and Polo (2003), including ex ante deterrence from spontaneous self-reporting, thereby bringing the analysis closer to the modern literature on optimal law enforcement, started by Becker (1968). In particular, Spagnolo (2000a, 2004) builds a model that focuses on general deterrence and the ability of leniency to induce undetected cartels' members to report their conspiracy, and shows that leniency programs open to firms reporting before an investigation can have a powerful deterrence effect not accompanied by prosecution costs. This follows from the model accounting for (1) the fact that firms that deviate from a cartel and report under leniency programs are protected by public antitrust enforcement afterwards instead of remaining liable for years (the "protection from fines" effect); and (2) the fact that a well-designed leniency policy that rewards with immunity the first reporting party and exposes its partners to harsh sanctions can generate general deterrence through the fear of being betrayed without any cost of prosecution (the "strategic risk" effect). Then, the model shows that: (i) having a leniency policy for firms reporting before an investigation is always optimal; (ii) these programs must be strict, that is, they must restrict (full) leniency to the first reporter only, so as not to shut down the two most crucial deterrence channels; (iii) if accompanied by severe enough sanctions, they could completely deter cartels – at a finite level of fines – even if the probability of detection of the cartel without a leniency application goes to zero; (iv) general deterrence is particularly powerful if rewards are introduced, funded by a fraction of the fines paid by firms that did not apply for leniency (so that the first-best scenario – costless and complete deterrence – could theoretically be achieved); and (v) the additional deterrence channel opened by well-run, strict leniency programs – with or without rewards – because they increase distrust/fear of being betrayed by other cartelists, can be captured by an extended notion of "riskiness" related in spirit to Harsanyi and Selten's (1998) concept of risk dominance. This novel deterrence channel adds to the deterrence channels linked to the participation constraint, identified by Becker (1968), and to the one linked to the cartel's incentive constraint identified by Stigler (1964), and is the crucial one to take into account when designing leniency programs, as it

becomes active earlier than the other, hence at much lower cost to society (as also confirmed by the experimental evidence discussed later).³

Spagnolo's (2004) model does not distinguish between colluding individuals and firms, as it focuses on the potential of these programs to fight many forms of collaborative/organized crime besides cartels, most of which involve multiple collaborating individuals, not necessarily firms. When colluding agents are firms, as is the case for cartels, and rewards can be paid to individual employees of these organizations, a number of novel issues emerge. Aubert et al. (2006) focus on the direct, general deterrence effects of leniency and rewards, and more precisely on those crucial "organizational" aspects. The model shows that (1) allowing whistleblowers of colluding firms to obtain leniency and cash a reward increases the number of potential informants that a colluding firm must "bribe" to keep silent, thus increasing the cost of collusion and, indirectly, the general deterrence effect of any given reward scheme; and (2) individual rewards tend to be complementary to corporate leniency programs, as they make a colluding firm's strategy to defect, report, and stop "bribing" its own informed employees even more attractive, further destabilizing collusion.

The authors also consider several potential costs from offering individual rewards, based on the possible negative effects on firms' internal organization and performance: deterrence of productive cooperation, which could be untruthfully reported in the attempt to cash a reward; inefficient reduction of the turnover to minimize the number of informed parties; and adoption of "innocent" attitudes, such as increasing investment in productivity-enhancing technology.

Potentially negative effects of leniency policies have also been noted. Motta and Polo (2003) focused on one negative effect of leniency: reduced overall cartel sanctions; Spagnolo (2000a, 2004) a second one: "exploitability" by repeat offenders (if the program is poorly designed). Buccirossi and Spagnolo (2006) highlight the importance of leniency policies for other forms of collaborative crimes like corruption and financial fraud, and identify a third negative side-effect of poorly designed programs. Focusing on leniency on bilateral, sequential, asymmetric illegal transactions, such as corruption or manager/auditor collusion, they show that "moderate" forms of leniency typically implemented in the real world could facilitate the enforcement of occasional illegal transactions by making the threat to report under the leniency program when one party violates the illegal agreement credible (see Spagnolo, 2000b for the case of multi-unit auctions). Related counterproductive side-effects of leniency are discussed in several other models (Ellis and Wilson, 2003; Brisset and Thomas, 2004; Motchenkova, 2005; Aubert et al., 2006; Harrington and Chen, 2006), confirming and reinforcing the message that leniency programs must be carefully designed and implemented, as they may otherwise produce rather negative side-effects. To close the "early studies" section, Harrington (2008) removes the assumption of a fixed and/or restricted⁴ probability of detection (and conviction) over time. In earlier papers, this assumption produced the counterfactual that, in equilibrium, colluding firms do not use the leniency program. In this study, the model setup includes a dynamic setting in which the probability of detection varies stochastically over time. The author shows that the overall effect of leniency

³ See Blonski, Ockenfels and Spagnolo (2011), Blonski and Spagnolo (2014), Bigoni, Casari, Skrzypacz and Spagnolo (2015), and Breitmayer (2015) for clarifications on the game-theoretic foundation of the strategic risk channel and robust experimental support for it.

⁴ Motta and Polo (2003) allow the probability of detection to change over time but it is restricted to only two values, one of which is zero. Feess and Walzl (2004) also allow this probability to be random but the setup is static and therefore, collusion is not endogenized.

programs depends on firms' incentives and the size of the fine reduction. In particular, a more generous fine reduction makes the cartel less stable, such that it is best to offer immunity to the first self-reporting firm only.⁵

2.2 More Recent Theories

The relationship between the leniency rate and the number of reporting firms is examined by Sauvagnat (2014) who extends Spagnolo's (2004) model to allow the competition authority to offer leniency rates contingent on the number of reporters. In this case, the optimal solution is not restricting leniency to the first reporter but instead only granting leniency if there is only one reporter. This is because this rule increases the expected sanctions of cartel members and thus, increases deterrence. Unfortunately, the model does not analyze the effects of these rules (single reporter or first reporter) on prices and, more importantly, on cartel formation.

Harrington and Chang (2009) design a dynamic model of cartel formation and dissolution, where an industry of firms interact repeatedly in an infinite time horizon, the population of cartels and of discovered cartels is endogenized and industry heterogeneity is captured by a stability parameter. At the start of each period, an industry may be cartelized or not, depending on whether it was cartelized in the previous period or not, and in the latter case there is uncertainty as to its cartelization in the current period. Firms realize profits and at the end of the period, a competition authority may arbitrarily start an investigation. If convicted, the severity of the fines depends on the number of collusion periods, given that there is only one collusive price in the model. This study shows how changes in the duration of cartels discovered can be instructive in assessing the effectiveness of the leniency policy in reducing the true rate of cartels. This is because if discovered cartels tend to be longer in duration following a policy change, then this change is likely to be reducing cartel formation.

While informative, the model assumes that there is no (threat of) entry or exit in each industry and the detection probability is exogenous. The latter issue is addressed in two extensions of this study: Harrington and Chang (2013) endogenize non-lenient enforcement and Harrington and Chang (2015) allow the competition authority to decide on its caseload.

The first extension shows that a leniency program is able to lower the cartel rate when the case requires sufficiently few resources or when enforcement was initially very weak. Otherwise, the program may actually increase the cartel rate. However, this negative effect can be mitigated if leniency programs are complemented by high penalties and a sufficiently large budget for competition authorities.

In the second extension, Harrington and Chang show that changes in the average duration of convicted cartels should follow a precise temporal pattern, after a policy is introduced. If the policy innovation is successful in increasing cartel deterrence, we should observe an increase in the average duration of convicted cartels in the short run. This is because less stable cartels, with lower expected duration, immediately disintegrate; ensuing cartel detections will therefore come from a population of more stable cartels, which typically last longer.

⁵ This statement holds if the cumulative distribution function of the probability of detection is weakly concave and the probability of a conviction, without self-reporting, is less than 50 percent.

2.3 Leniency With Asymmetric Firms

Recently, some authors have formally examined the effect of asymmetric firms and emphasized that the deterrent effect of leniency programs depends, to some extent, on the heterogeneity of the cartel members. In this sub-section, we discuss articles that consider heterogeneous firms in terms of (1) their size (Motchenkova and Van der Laan, 2011; Motchenkova and Leliefeld, 2010); (2) their role as instigator and/or leader of the cartel (Herre, Mimra and Rasch, 2012; Bos and Wandschneider, 2012; Chen, Ghosh and Ross, 2015; Blatter, Emond and Sticher, 2014); or (iii) the possession of private information (Feess and Walzl, 2010; Silbye, 2010; Pinna, 2010; Harrington, 2013; Marvão, 2014).

Motchenkova and Van der Laan (2011) extend the model by Motta and Polo (2003) to include firms that collude in one market but differ in their diversification in other non-cartelized markets. Therefore, the novelty in the model setup is that firms are of different sizes and operate in several markets, but collude in (only) one of them. This setup generates an additional cost in the case of reporting of the cartel, due to the possibility of asymmetric punishments, that is, punishing the reporter in a different market.

In this study, leniency programs are shown to be more effective for smaller firms (where inducing self-reporting requires a lower rate of law enforcement), and require larger fine discounts where resources are limited, as shown in previous work. In addition, the model suggests that larger firms prefer not to enter a collusive agreement than to collude and report and so, the competition authority is more able to prevent them from collusive activities.

A different source of heterogeneity is examined by Motchenkova and Leliefeld (2010). The authors use a similar structure to Houba, Motchenkova and Wen (2011) but in their duopoly model, firms have different accumulated profits (i.e., market shares). In this setting, a small firm decides whether or not to report the cartel, after which a large firm decides whether to retaliate or not. As such, the large firm can use the leniency program to coerce and prevent the small firm from reporting. Those with higher accumulated profits are able to prevent the other (smaller) cartel members from reporting, through predation. In this case, leniency is not effective, as it strengthens cartel stability and leads to an abuse of a dominant position. As previously shown in Ellis and Wilson (2003), the results emphasize that leniency programs may serve as a disciplining device to hinder defections from the collusive agreement, unless reporting firms are protected from possible retaliation by the other cartel members.

Both these studies assume that all firms are equally efficient (same production cost); the firm asymmetry does not allow for bargaining power within the collusive agreement; and both collusive prices and profits are fixed. Removing these assumptions would make these models more realistic. Even more, the reasons underlying the heterogeneity in firm size are not explored. One possibility is that firm size is related to being the cartel instigator and/or leader.

The US “Corporate Leniency Policy” confines leniency applications to firms that “did not coerce another party to participate in the illegal activity and clearly was not the leader in, or originator of, the activity”.⁶ However, some jurisdictions (the EU in 2006 and Canada in 2010) have removed such a restriction.

⁶ US DOJ, Antitrust Division, “Corporate Leniency Policy” at A.6, available at <http://www.justice.gov/atr/public/guidelines/0091.htm>.

To compare leniency regimes where the ringleader is excluded or not, Herre et al. (2012) extend the model by Motta and Polo (2003) in which firms set quantities and then decide to report or not, and the competition authority secures a conviction, as long as a cartel member reports. In this study, each cartel member has some evidence that affects the probability of conviction but no firm has sufficient evidence to guarantee a conviction. In addition, side-payments between cartel members are permitted. The authors show that the introduction of a “no immunity to instigator (or ringleader) clause” has little effect if the instigator has a large amount of evidence to provide authorities with, particularly if the probability of an investigation is low. This is because the dominant collusive agreement between the members specifies that the ringleader never reports and is compensated for its silence. However, if the ringleader’s information is unimportant, the results show that he should be excluded from leniency.

Other authors have found a less subtle effect from excluding ringleaders. Motta and Polo’s (2003) model is also extended by Chen et al. (2015) and Blatter et al. (2014). The first study adds an instigation stage whereby a firm may elect to suggest a collusive agreement, and a special treatment of instigators whose exclusion from the leniency program eliminates their incentive to report. If both firms simultaneously decide to collude, one is randomly assigned the role of instigator. In the second study, the model is extended so as to consider the minimum standard of evidence and instigator discrimination, such that firms have asymmetric and imperfect cumulative evidence of the collusion.

Chen et al. (2015) formally show the dual effect on firms’ incentives. Excluding instigators from leniency decreases the destabilizing effect of leniency programs on cartels, thus increasing cartel stability. However, by imposing a more severe punishment on the instigator, it may reduce the incentive for cartel instigation. These results hold for perfect and symmetric information. Blatter et al. (2014) corroborate the results from Herre et al.’s (2012) study and add the result that asymmetric evidence increases the cost of deterrence if the firm with most evidence is excluded from the program and thus does not report.

Bos and Wandschneider (2012) add to this result. The authors extend the price-setting supergame with capacity constraints framework from Bos and Harrington (2010) and endogenize the composition of a cartel in the context of an infinitely repeated game with heterogeneous firms. The collusive price levels depending on the exclusion or not of the ringleader are analyzed and it is shown that introducing such a clause will generally (although not always) lead to lower cartel prices.

Silbye (2010) explores the issue of a two-firm cartel where firms have imperfect and asymmetric evidence.⁷ In his framework, two firms possess different amounts of evidence but only one is allowed to apply for leniency, and the competition authority sets a fine that decreases with a larger amount of provided evidence. The author assumes that the likelihood of detection of the cartel is common knowledge but each firm has evidence that it could submit to convict the other firm if it applied for leniency. The findings are in line with the current EU leniency program: firms who provide more evidence should receive a larger leniency reduction, and evidence should be significant and relevant in order to lead to a larger reduction, provided that fines are not too low.

⁷ See also Feess and Walzl (2010) where a similar issue is addressed but in a static model of collusion. Silbye’s (2010) analysis extends their analysis to a setting with repeated interactions.

In Harrington (2013), each cartel member has private signals on the probability of being detected by a random audit and will only apply for leniency if its signal is above a given threshold. The model examines the effect of leniency in this scenario and it emphasizes a problem that is denoted the “pre-emption effect”, in the sense that a firm may apply for leniency because it fears another firm will apply. This channel is related to the strategic risk channel analyzed in Spagnolo (2004) but the introduction of private information means that the fear of being cheated upon by a cartel partner may as well induce reporting along the equilibrium path, which is very important for connecting to empirical studies.

The assumption of symmetric firms in Harrington (2013) was later removed by Marvão (2014), who shows that the optimal reporting strategy of cartel members is no longer symmetric but it is the firm that expects to receive the largest fine (due to larger sales, recidivism or other aggravating circumstances) that reports the cartel first.

Sauvagnat (2015) examined the role of private information. The novelty in this infinitely repeated Bertrand game is that the competition authority has private information on the strength of a case, and it is a strategic decision to open a case. This leads to the result that leniency should be granted even when the probability of conviction is high, as it increases the rate of success of the cartel investigation. Conversely, the authority is also able to exploit firms’ uncertainty regarding the likelihood of conviction and it may open an investigation even when its strength is low, as that may induce firms to apply for leniency.

2.4 Leniency and Investigative Effort by Competition Authorities

Chen and Rey (2013) analyze the trade-off between leniency programs and the effectiveness of antitrust investigations. The authors use a two-firm infinitely repeated Bertrand game in the spirit of Spagnolo (2004) to study when it is beneficial to allow leniency to be also awarded after a cartel is detected and an investigation started, and when instead it is better to only offer leniency before an investigation is started. As in Spagnolo (2004), and in contrast to Motta and Polo (2003) they show that leniency should be restricted to the first reporting firm (regardless of recidivism); that it is always beneficial to offer leniency before an investigation started; and imply that rewards for the first spontaneously reporting party may be optimal. Their most novel contribution is the full characterization of the conditions under which it is also optimal for society to allow competition authorities to offer some leniency after a cartel is detected and an investigation has started, as done in reality. They show that in the case where investigations are infrequent but likely to succeed, it is optimal to offer less leniency once an investigation is ongoing; whereas when investigations are frequent but unlikely to succeed, it may be optimal to offer more amnesty once an investigation is initiated, so as to increase its effectiveness. The authors also compare different characteristics of leniency programs, providing some support for the “first informant” and the “post-investigation amnesty” rules that are present in the US leniency program. However, no support is found for not offering leniency to repeat offending firms.

Gärtner (2013) extends this framework to include the feature of memory in the stochastic process, that is, the probabilities of detection evolving over time. Adding this persistence feature pushes firms into pre-emptive leniency applications, which occur much before the risk of independent detection is impending. This pre-emptive effect is larger if there is little discontinuity in time and state, low level of firms’ patience, and a relatively severe punishment

of firms that fail to pre-empt others. These results hold even in the absence of rewards or large absolute levels of leniency reductions.

The literature above assumed that competition authorities are able to commit to an investigation effort *ex ante* (or else that the probability of a conviction is exogenous). Gerlach (2013) removes this assumption and develops a model where the authority has two instruments: a self-reporting scheme and procedural investigation. An instrument is chosen in the first game stage. In the second stage, the amount of private information of the wrongdoer, which is common knowledge, is randomly assigned and the wrongdoer then decides to commit the crime or not. If it does, then it decides to report it or not. In the following stage, if there is no report, the authority chooses its effort level and convicts or acquits the wrongdoer.

The model shows that the lack of commitment generates a negative relationship between investigative effort and self-reporting schemes, such that self-reporting is more efficient when: (1) full amnesty is offered, as wrongdoers (always) report; (2) the level of harm is low; and (3) authorities can convict without hard evidence, although a hard-evidence standard provides more deterrence (and is welfare enhancing).⁸

2.5 Other Issues Related to Leniency Programs: Multi-market Contact, Price Effects, Damages and Delegation

Few papers have examined the specific case of leniency applications by multi-market firms. Choi and Gerlach (2012) are the first to examine multi-market contact in the context of cartels. In order to do so, the authors extend the framework from Motta and Polo (2003) to allow for collusion and self-reporting where firms form cartels in two geographical markets (i.e., countries) and national competition authorities may or not cooperate between themselves. If there is only one authority per market, free-rider problems arise due to positive prosecution externalities in each market. These can be solved if competition authorities are able to cooperate. However, if they share extensive (relevant) information, the incentive to self-report is decreased. Conversely, if there is no cooperation, multi-market contact allows firms to reduce the equilibrium amount of self-reporting and sustain collusion more effectively.

Under the US “amnesty plus” program, if a firm is convicted in a market, it is asked if it colludes in other markets and if it does not report these cartel(s), it can no longer apply for leniency if an investigation is open on those cartels. Unfortunately, the paper above is not able to capture the strategic effect of leniency or amnesty plus programs. Lefouili and Roux (2012) do this using a dynamic framework to show that amnesty plus affects collusive strategies in which firms cooperate and never report the existing cartels. The program is shown to be able to destabilize cartels where firms continue to collude after the detection (and conviction) of the first cartel (pro-competitive effect). However, it is also able to stabilize cartels where firms reveal the second cartel after prosecution of the first cartel (pro-collusive effect).

Marx, Mezzetti and Marshall (2015) add to this literature by adopting a static model that allows for the use of global games techniques to capture mis-coordination problems, rather than a dynamic model with self-enforcing constraints. They focus on multi-product

⁸ The effect of judicial errors in the presence of a leniency program has been analyzed by Motchenkova and Ghebrihiwet (2010) and Pavlova and Shastitko (2014), showing that judicial errors (types I and II) make antitrust enforcement less effective and *ex ante* deterrence weaker.

colluders and examine their incentive to report the different cartels they are involved in, under the US amnesty plus program. They show that it is possible that linking leniency across products increases the likelihood of conviction in the first product investigated but reduces it in subsequent products. Thus, firms may have an incentive to form sacrificial cartels and apply for leniency in less valuable products to reduce convictions in more valuable ones.

Houba, Motchenkova and Wen (2015) study the impact of the cartel destabilizing effect of leniency programs on maximal pricing. The authors develop an infinitely repeated oligopoly model where fines are a function of the illegal gains and where the probability of detection depends on the degree of collusion, and that focuses on the worst possible scenario. To solve this complex model the authors have to assume, as in Motta and Polo (2003), that if a firm deviates, it immediately becomes no longer liable to antitrust fines (in most jurisdictions, instead, a cartel member that deviates and abandons a cartel remains fully liable to public and private enforcement for several years). This simplifying assumption, as explained earlier, is not innocuous as it shuts down entirely the “protection from fines” (deviating, and deviating and reporting then have the same effect) and the “strategic risk/fear of betrayal” effects of leniency, crucial according to experimental evidence. Not surprisingly, assuming away these effects, the model misleadingly predicts, as in Motta and Polo (2003), that leniency before investigation can have no effects. More interestingly, when studying post-investigation leniency, the authors find that strict programs offering full immunity only to the first reporting firms are the best scenario in terms of the size of the reduction in the maximal cartel price. This last result, that leniency programs should be strict, appears likely to be robust to more realistic assumptions and therefore policy relevant.

Public and private enforcement typically serve complementary purposes. However, modern antitrust engenders a possible conflict between the two due to the central role of leniency programs. Damage actions may reduce the attractiveness of leniency programs for cartel participants if their cooperation with the competition authority increases the chance that the cartel's victims will bring a successful suit. Buccirossi, Marvão and Spagnolo (2015) examine the EU Directive on Antitrust Damages Actions, adopted in November 2014, which seeks a balance between public and private enforcement. For this purpose, they develop a simple model where in addition to fines, cartel members are liable for compensation of buyers. The analysis shows that damage actions are able to improve the effectiveness of leniency programs through a legal regime in which the civil liability of the immunity recipient is minimized and full access to all the evidence collected by the competition authority, including leniency statements, is granted to claimants.

Earlier research (Spagnolo, 2000a, 2005) empirically observed that incentives to top managers, such as staggered stock options and bonus schemes, tend to facilitate collusive behavior. Chen (2008) adds to this strand of literature by characterizing the effectiveness of leniency programs for deterrence, in centralized and decentralized cartel hierarchical organizations. The author shows that incorporating non-contractibility issues may further facilitate collusion. However, the efficiency gains of delegation in facilitating collusion can be mitigated when the leniency program is introduced, particularly when the probability of detection, absent leniency, is low and corporate liability is much more significant than the individual one.

Angelucci and Hann (2011) build a three-tier model where an antitrust authority, a shareholder and a manager interact, in order to study a firm's internal agency problem. The shareholder monitors the manager's behavior and although the shareholder is not directly

affected by the manager's behavior, both can report the infringement but it is assumed that the manager is quicker to report than the shareholder. If neither reports, the authority can still uncover the cartel with some probability. The authors show that it is optimal to grant a partial reduction of the corporate sanction, in exchange for evidence of the misbehavior, but to fully punish the manager. Individual leniency programs are also shown to increase private compliance costs, that is, transferring part of the cost to the shareholder, and should be offered to a reporting manager when the ability of the authority to punish management is limited.

Dargaud and Jacques (2016) analyze the possibility of using leniency programs (and amnesty plus) to induce CEOs of decentralized firms to carry out internal audits and report collusive agreements. The authors develop a simple model where two homogeneous firms produce two different products and play an infinitely repeated Bertrand game, in the presence of an antitrust authority. Two scenarios are illustrated, where firms are able or not, time-wise, to report a second cartel before its detection. If they are not able to report it, in the presence of leniency programs (but not amnesty plus), CEOs will not investigate a potential second cartel and report it but the sustainability of the cartel is decreased by the presence of leniency. In the second scenario, where reporting is feasible, CEOs may start an investigation if the level of compartmentalization is not perfect. In the presence of amnesty plus, the results show that there is a pro-collusive effect for centralized firms who simultaneously collude in two markets.

3 EMPIRICAL EVIDENCE ON LENIENCY

The number of different models with often unrealistic simplifying assumptions and conflicting results calls loudly for empirical and experimental evidence. Let us now turn to survey this.

3.1 Descriptive Studies

The empirical literature on leniency policies is recent and includes several papers that present and discuss descriptive statistics on prosecuted cartels (Bloom, 2007; Connor, 2007, 2013; Carree, Günster and Schinkel, 2010; Combe and Monnier, 2011; Veljanovski, 2010; Dominte, Ţerban and Dima, 2013). A case study approach has also been used to examine leniency policies (Asker, 2010).⁹

A yearly analysis of leniency applications in the EU and the USA clearly shows that the number of cartels reported under a leniency policy and the number of individual leniency applications have both increased dramatically in recent years. This is particularly the case in the EU. The generosity of the penalty reductions for initial and subsequent leniency applicants in the EU has also visibly increased (see Marvão, 2014).

Between 1996 and March 2010, 124 firms were fined by the DOJ for participation in 39 different cartels.¹⁰ In the EU, leniency applications in the period between 1998 and October 2014 related to 81 cartels, with a total of 385 firms. In Asia, 33 cartels were prosecuted by the Japan Fair Trade Commission, the Korea Fair Trade Commission and Taiwan's Fair Trade

⁹ Asker has made an in-depth analysis of a US parcel tanker shipping cartel and suggested that welfare-improving effects from leniency are linked to an increased probability of private antitrust suits.

¹⁰ See further statistics for the USA in Connor and Miller (2013); for Asia in Connor (2007); and for Russia in Yusupova (2013).

Commission between 1990 and 2007. Finally, 30 cartels were prosecuted in Russia between 2004 and 2011. Furthermore, the average number of cartel members in the reported cartels is smallest in the EU (8.3) and largest in Russia (11), although the latter figure is inflated by a cartel involving 51 firms in the financial services industry between 2003 and 2008.

Repeat offenders are a highly debated issue (see Werden, Hammond and Barnett, 2011). Connor (2010) has suggested that there is evidence of a large amount of recidivism; he identified 389 recidivists worldwide in the period between 1990 and 2009. This number constitutes 18.4 percent of the total number of firms involved in 648 international hard-core cartel investigations and/or convictions. Werden et al. (2011), however, have contested Connor's definition of recidivism and his calculation of the numbers of multiple and repeat offenders. The main discrepancy between the two arguments appears to be in the examined period and in how cartel members who merge and form a new firm are dealt with. Werden et al. follow the legal practice (DOJ and EC) and therefore, they have suggested that no repeat offenders in US cartels have been fined since 1999. As for the EU, Marvão (2015) identified 89 multiple offenders and six repeat offenders¹¹ since 1998 when the first leniency reduction was granted. The first decision applying the leniency policy to a cartel case was in 1998, involving British Sugar.¹² The complaint was made in 1994 and, after the introduction of the leniency policy, all four cartel members applied for leniency. Three reductions of 10 percent and one of 50 percent were granted.

On average, a cartel member fined by the EC receives a leniency reduction of 26 percent. Firms that receive a leniency reduction (1–99 percent) receive first, on average, a fine increase of 32 percent (these are granted for aggravating circumstances such as recidivism, absence of cooperation, obstruction of the investigation and for being the cartel ringleader or instigator) and a fine reduction of 3 percent (for mitigating circumstances such as termination of the infringement at the time of the investigation, negligence as the cause of the cartel, limited involvement in the cartel, cooperation with the Commission outside the leniency policy, or proof of having been encouraged by public authorities or legislation).¹³ The average firm that does not receive a leniency reduction has a fine reduction of 16 percent and an increase of 56 percent, whereas firms with immunity from fines would have faced, on average, a fine increase of 22 percent and a decrease of 3 percent.

Recently, Uytsel (2015) offered a first analysis of leniency in Japan, introduced in 2005 under the Japanese Antimonopoly Law. The author describes the 236 leniency applications in Japan between 2006 and 2013, where 28 percent of the cartel members received immunity from fines. All the leniency applicants are domestic firms (even those in international cartels), most of which are listed. This is in line with empirical studies showing that the first leniency applicant is the largest firm (see, e.g., Marvão, 2014). Uytsel also presents the results of a survey that was sent to leniency applicants in the years of 2006 and 2012, although the response rate was very small (15/80). The article raises doubts as to the effectiveness of the leniency program, suggesting that too many unmeritorious applications reach the Japan Fair

¹¹ These are defined as any firm who was caught colluding after having received a fine for another cartel. In this sense, the definition is closer to Werden's than to Connor's.

¹² For information on the case see: http://ec.europa.eu/competition/antitrust/cases/dec_docs/33708/33708_6_7.pdf.

¹³ The fine is capped at 10 percent of the total turnover of the firm in the previous year. Special conditions are set in the case of inability to pay. For further information on cartel fines, see guidelines on the method of setting fines imposed pursuant to Article 23(2)(a) of Regulation No 1/2003 [2006] OJ C210/2.

Trade Commission and that too many firms, per cartel, receive leniency reductions. Even more recently, Wils (2016) assembled a rich descriptive analysis of the overall European experience with leniency programs, from its start to 2015.

Although descriptive statistics (and case studies) are important to show correlations and trends, they fail to identify causality and therefore cannot be used to evaluate the consequences of a policy, such as the deterrent effect of a leniency program. The real impact of leniency policies can only be addressed with econometric methods, but methodological problems (such as potential sample selection bias from only observing detected cartels) and the lack of appropriate data make empirical research scarce. The results of econometric studies, considered next, must therefore be interpreted with some caution.

3.2 Econometric Studies

3.2.1 Deterrence effects of leniency policies

As we suggested earlier in this chapter, the most important effect of leniency policies is the resulting (hopeful) decrease in the number of cartels in society. However, this is very difficult to measure because only *detected* cartels are typically observed. Two methodologies have been developed to infer the effects on cartel formation and deterrence of changing a law enforcement policy.

Harrington and Chang (2015) studied a dynamic model of cartel formation and showed that changes in the average duration of convicted cartels should follow a precise temporal pattern, as described in Section 2.2 of this chapter. The second methodology was derived by Miller (2009), who developed a somewhat simpler dynamic model of cartel behavior from which he derived predictions for successful law enforcement innovations related to the temporal distribution of the number of detected cartels conditional on the leniency policy. His model suggested that: (1) an immediate increase in the number of detected cartels is consistent with the hypothesis that a leniency policy increases the probability of cartel detection; and (2) a subsequent decrease in the cartel detection rate and stabilization at a constant level lower than the one prevailing before the introduction of the leniency policy is consistent with it having a significant deterrence effect, that is, with the policy reducing cartel formation. The first and only two studies we are aware of that apply these methodologies to cartel data are studies by Brenner (2009) and Miller himself.

Brenner studied EU cartel cases in the period between 1990 and 2003, a dataset that included 61 cases. He tested the evolution of the average duration and number of cartels detected around and after the introduction of the first version of the EU leniency policy in 1996.¹⁴ He found neither an increase in average duration after the introduction of the leniency policy nor an increase in the number of detected cartels immediately after the policy's introduction. And he did not find a decrease in the number of detected cartels in the longer run. These findings appear inconsistent with the theoretical conditions indicating that the 1996 leniency policy had positive deterrence effects.

While Brenner's conclusions are consistent with the general perception that the 1996 EU leniency policy was rather poorly designed and implemented (see also Bloom 2007) (it was

¹⁴ See Commission Notice on the non-imposition or reduction of fines in cartel cases [1996] OJ C207/4-6.

reformed in 2002¹⁵ and again in 2006¹⁶), it would be interesting to subject his findings to a number of robustness checks. For example, he treated the first three years of the leniency policy's existence as the short run, but there is no clear definition from Miller's or Chang and Harrington's studies as to how the short run should be defined. Hence, one would want to see analogous tests for a large number of other time frames to be confident about robustness.

Miller applied his own methodology to assess the effect of the reformed US leniency policy introduced in 1993.¹⁷ He used data from the US DOJ that cover the period between 1985 and 2005. He found that the number of cartels detected by US authorities increased after the introduction of the new leniency policy, which according to his theory is consistent with an increase in the cartel detection rate. He also observed that this increase was followed by a fall to a level below the pre-lenency policy level, a pattern that according to his theory is consistent with increased cartel deterrence. The mentioned changes in the number of detected cartels were statistically significant, of a large magnitude and consistent with several robustness checks.

Although Miller's study probably represents the most important contribution to the empirical literature on the effects of leniency policies to date, it has not escaped criticism. Cartel formation and dissolution are not endogenized in the model, although this seemed to be present in an earlier draft of the paper. Another issue is that changes in the cartel duration of detected cartels were not considered as a robustness check.

De (2010) also tested Harrington and Chang's theory. The paper focused on the precise determination of the life-span of 109 EU cartels that were the subject of an infringement decision between 1990 and 2008. Previous empirical studies on cartel duration used methods that assumed a normal distribution of the lifetime data, and they were unable to deal with a flexible probability of exit from a cartel or with more than one reason for a cartel breaking up. To overcome both of these issues, De analyzed the dataset with the help of a competing risk Cox proportional hazard model. The regression results showed that the introduction of a leniency policy was one of the causes of cartel breakdown. De argued that it is extremely difficult to empirically define the short and long run and so her model refrains from doing so. Nonetheless, her results showed that cartels detected after the introduction of the initial leniency policy in 1996 had a lower survival probability than those cartels that were detected earlier. According to Harrington and Chang's model, this finding is not consistent with an increase in deterrence linked to the leniency policy.

Zhou (2013) applied Harrington and Chang's model to EU leniency policy data for the period between December 1985 and December 2011. Using hazard model regression estimates, he found that cartel durations increased significantly in the period immediately following the introduction of a leniency policy (consistent with enhanced detection) and subsequently fell below short-run levels (consistent with enhanced deterrence). In addition to providing results supportive of the 2002 EU leniency policy, Zhou's paper also tried to improve methodologically on previous work. He argued that De did not differentiate the short-run from the long-run impacts and that Brenner (2009) took the first three years of the leniency policy's existence as the short run without theoretical support for doing so. Zhou differentiated

¹⁵ See Commission Notice on immunity from fines and reduction of fines in cartel cases [2002] OJ C45/3-5.

¹⁶ See Commission Notice on immunity from fines and reduction of fines in cartel cases [2006] OJ C298/17.

¹⁷ See US DOJ, "Corporate Leniency Policy" (10 August 1993) at www.justice.gov/atr/public/guidelines/0091.pdf.

the impacts by cartel start date, which is more in line with Harrington and Chang's model: the short-run impact arises only with cartels that started before the introduction of the leniency policy, and the long-run impact arises only with cartels born after its introduction.

Klein (2011) tried to identify the deterrence effect of leniency policies by directly linking their introduction to an indicator of competition intensity. His empirical analysis relied on Organisation for Economic Co-operation and Development data for the period between 1990 and 2010 and included 23 countries. He calculated the average profitability of industries (quotient between value added and cost of capital and labor), which he then used to draw inferences about the price–cost margin, since both are directly related. Issues of sample selection bias, endogeneity and omitted variables were addressed through the use of additional control variables (changes in GDP trend, in imports and in import penetration), an instrumental variable estimation and several robustness tests. The results showed that leniency policies were associated with a decrease in the price–cost margin of 3–5 percent. Unfortunately, the interpretation of the average profitability of industries is multi-faceted and its correlation with competition intensity is not entirely clear.

The instrumental variables approach used in Klein (2011) is criticized by Cloutier (2011), who argues that the estimates are biased due to endogeneity issues. Cloutier argues that for large supra-EU firms, leniency programs in other countries may affect the cartel profitability and the policy position of a country's political parties, making this an inadequate instrument. The author attempts to tackle the endogeneity concerns by using industry concentration as a proxy for industry competitiveness in a difference-in-differences regression, where the control group is low concentration industries. The results, which are consistent with Klein's, show that, for the USA, between 1991 and 1997, the amendment of the 1993 Corporate Leniency Policy had no significant short-run effect on price–cost margins but had a persistent effect after one to two years.

Yusopova (2013) has presented the first econometric assessment of the leniency policy that was introduced in Russia in 2007.¹⁸ The perceived ineffectiveness of the leniency policy led to a reform in 2009, when full immunity and criminal liability were introduced.¹⁹ The data included all 30 cartels that were fined between 2004 and 2011. There were up to 51 firms per cartel. The results from a Poisson regression showed that the 2009 revision of the leniency policy was associated with a decrease in both the size of detected cartels and their duration. The results also showed that industries²⁰ with low concentration have had fewer cartel convictions since the leniency policy has been in place. Yusopova concluded that the 2009 revision was effective. Harrington and Chang's theory would suggest the opposite, however: a decrease in the duration of detected cartels is consistent with the new policy causing a reduction in cartel deterrence.

Dong, Massa and Žaldokas (2014) use data on nearly 489 000 registered global firms, over the period of 1990 to 2012, to study the impact of leniency laws on firms' strategies and

¹⁸ The leniency policy was established by Federal Law No. 45 "On Amendments to the Russian Federation Code of Administrative Offences". A previous discussion of the Russian leniency program is carried in Shastikov and Avdasheva (2011).

¹⁹ See Code of the Russian Federation on Administrative Violations, Art. 14.32 notes (civil liability); Criminal Code, Art. 178, Note 3 (criminal liability). See generally Federal Antimonopoly Service of the Russian Federation, at [http://en.fas.gov.ru/upload/other/International%20Cartel%20Investigations%20-%20Main%20Steps%20\(M.%20Khamukov\).pdf](http://en.fas.gov.ru/upload/other/International%20Cartel%20Investigations%20-%20Main%20Steps%20(M.%20Khamukov).pdf); and Yeremin, Subbot and Mouradov (2011).

²⁰ The seven industries are defined according to the cartel reports from the Federal Antimonopoly Service and the concentration of each industry is categorized into high, medium or low.

the overall cost of collusion, using a difference-in-differences approach. As in most previous studies, it is shown that the introduction and existence of leniency policies increase the number of firms and cartels convicted. However, this study also finds that these policies are associated with a decrease of the gross margin of firms (not necessarily cartel members) by six percentage points, and that this is particularly strong in industries where collusion tends to be less stable. While these results suggest that leniency policies are effective, the study finds that after their introduction, firms pursue more mergers and acquisitions with firms in the same industry. That is, horizontal mergers seem to counterbalance the negative effect of leniency on prices, particularly when the acquirer had already been convicted for collusion.

Davies, Ormosi and Graffenberger (2014) and Marx and Zhou (2014) also examine this dynamic of mergers on detected cartels. The first authors restrict their sample to post-collusion periods (84 EC decisions taken between 1984 and 2009), to focus on the firms' post-dissolution activity, while the latter expands the sample to mergers that took place before the dissolution of the cartel (151 EC decisions taken between 1985 and 2013). To identify the impact of the leniency program, Davies et al. use the cause of the investigation in a continuous time method whereas Marx and Zhou use the date of the policy introduction with a discrete-time method, to tackle issues of reverse causality and measurement.

The results from the recurrent event survival analysis in Davies et al. show that mergers are more frequent after the cartel breakdown, particularly in less concentrated markets. However, the authors are not able to disentangle coordinated effects mergers (see Fabra and Motta, Chapter 5 in this volume) from those caused by market restructuring and that may be pro-competitive and thus, the effect of leniency is not well established.

Marx and Zhou are able to test how the leniency program affects post-cartel merger activity. This is done by using a reduced-form Poisson regression to test merger rates following leniency introduction (or a settlement procedure) and a discrete-time hazard regression to test if leniency is able to expedite mergers (and if settlements delay mergers). The model shows that the EC leniency program expedites mergers (while settlements delay them) and mergers occur at a faster rate post-cartel, in line with the results found by Dong et al. (2014).

Finally, Bos et al. (2016) study the impact of competition enforcement on cartels. A theoretical model is developed, where firms produce differentiated products, at a given level of overcharges, and compete in an infinitely repeated Bertrand game. The model predicts that effective anti-cartel enforcement deters cartels with low overcharges and leads to lower collusive prices in the cartels that do form. This suggests that effective cartel policies should translate into fewer cartels with low overcharges and fewer cartels with high overcharges. This hypothesis is tested using John Connor's Private International Cartels (PIC) dataset, which includes data on 1500 overcharges for 500 cartels, and the authors use a quantile regression to compare the distributions of legal and illegal overcharges. The empirical results corroborate the theoretical prediction and show that illegal cartels are less likely to set either low or high overcharges, providing some evidence of a deterrent effect from competition policies.²¹

²¹ Marvão and Spagnolo (in progress, September 2017) apply this methodology to a similar cartel overcharge dataset to test how deterrence changed with the introduction of the US and EU leniency programs, with mixed (preliminary) results.

3.2.2 Other issues related to leniency policies

In his 2009 study, Brenner also estimated the factors that influence the absolute amount of the fine (and the fine reduction) and the duration of the investigation. Using ordinary least squares estimations, Brenner showed that the leniency policy increased the average reduced and total fines by around €16.5 million and €30.9 million respectively. Furthermore, the introduction of the leniency policy decreased the average duration of cartel investigations by around 1.48 years. The duration of the cartel and the number of firms and countries involved in each seemed to play no role in determining the fine, the fine reduction and the duration of the investigation. However, the number of cartel members presented a negative coefficient in the model for investigation duration.

While these extra results advance our knowledge on the effects of the 1996 EU leniency policy, Brenner's analysis could be improved by using a data deflation process for the absolute amount of the fine and by weighting the absolute value of the fines with the turnover of the firms in the cartel.

A later paper by Brenner (2011) examines the resource advantage of leniency applicants. Using the same dataset, the author uses a logit model to establish the differences in the decision to report, between large multinational and other firms. The results show that the former are more likely to report and cooperate with an investigation but no other characteristics of reporting and cooperating firms are identified as being significant.

A master's thesis by Arlman (2005) presented a second analysis of the EU leniency policy of 1996, using a dataset of 67 cartel cases convicted by the EC between 1990 and 2004. Arlman found that the leniency policy, measured by a dummy for whether or not a firm received maximum leniency, is positively correlated with the number of words in a decision (a proxy for the amount of information in the Commission's possession) and the gravity of the infringement, and negatively correlated with the duration of the investigation. In line with Brenner's finding, the paid fine was also found to be higher once the leniency policy was introduced, although Arlman measured the paid fine as a share of the firm's turnover, which is problematic because it creates a bias between more and less diversified firms. Given these results, the author concluded that the leniency policy was moderately effective.

Gärtner and Zhou (2012) focused on the delay with which a cartel is reported relative to the time of collapse of the cartel. They analyzed 96 EU cartel cases, of which 78 included leniency applications. Between July 1996 and 2006, 40 percent of the leniency policy applications experienced delays, often longer than ten months, relative to the time of collapse of the cartel. A hazard model, where spells correspond to periods of application delay, was used in the analysis of the leniency application. They found that the introduction of the EU leniency policy in 2002 had a negative effect on the decision to apply for leniency. Delayed leniency applications were also shown to be correlated with the severity of the punishment and with business cycles. These results were corroborated by probit model estimates and robustness checks.

Zhou (2016) adds to these findings by analyzing how firms' incentives for leniency applications change over time and how they are affected by different enforcement and institutional parameters. The author analyzes EC data between October 1996 and December 2014 through a multiple-spell discrete-time hazard regression. The main result of the study is that the start of an EC investigation does not affect the rate at which cartel members apply for leniency in the investigation market, but it does increase the rate of leniency applications in other markets in which one or more cartel members also engage in collusive activities.

Marvão (2015) has provided a more recent assessment of the EU leniency policy by examining the factors that encourage cartel members to self-report. The self-compiled data employed in the empirical analysis included all cartels up to October 2014 where there was at least one successful leniency policy application (87 cartels involving 510 firms). The study distinguished firms that started to participate in a cartel after being previously fined (four), those that ended their collusive behavior after being fined for participating in another cartel (six) and firms that ended their participation in a cartel after being investigated for a second cartel of which they were a participant (22). A total of 89 firms participated, contemporaneously or not, in at least two cartels (that is, they were multiple offenders). The econometric analysis, using double-sided Tobit and Heckman two-stage models, showed that the first reporter received much higher fine reductions, whether or not the reporting of the cartel took place before the EC started an investigation. The predicted leniency reductions were also larger for firms in smaller cartels, in cartels with a wide geographical impact and for firms that received lower fine reductions outside the leniency policy. The main result of this study is that repeat offenders appeared to receive higher leniency reductions, which suggests that firms can learn the “rules of the game”, repeatedly colluding and reporting the cartel, and thus substantially damage their partners.

In an earlier paper, Marvão (2014) studied the characteristics of the firms reporting under the leniency policy and the cartels they take part in. Probit estimates were carried out using self-collected EU data as in the earlier study (Marvão, 2015) together with US data from John Connor's PIC database. In the USA, in the period between 1984 and 2009, 2310 firms were convicted for their participation in cartel activities. The empirical analysis showed that EU firms that report the cartel and receive immunity from fines under the leniency policy are typically repeat or multiple offenders and are less likely to have received other fine reductions, while in the USA the reporting firms are more likely to be the cartel leader as defined in Connor's database.²² Repeat offenders were also more likely to receive immunity if they report once the collusive agreement ended. In contrast, firms that received other reductions were less likely to apply for and be granted immunity if the cartel is over.²³

Some of the characteristics of the cartels in which pre-investigation reporting occurs were also unveiled. In the EU, these cartels tended to be smaller in terms of the number of members (and also number of repeat and multiple offenders) and tended to impact a geographical area wider than the European Economic Area. Reporting was also more likely to occur in the fine art auctions sector, which has a small number of firms and where reporting will significantly damage the competitors that also took part in the cartel. In the USA, the predicted probability of immunity was much larger in the rubber and plastic sector and the paper and printing sector, and in markets with a moderate number of buyers.

On the issue of ringleaders, Davies and De (2013) empirically examined the frequency and characteristics of ringleaders in the EU and how they were treated when a leniency policy was introduced in 1996. Ringleaders were identified in one-fifth of 89 EU cartels

²² This result contrasts with the US leniency policy's statement that ringleaders cannot receive leniency (see US DOJ, “Corporate Leniency Policy”, 10 August 1993, at www.justice.gov/atr/public/guidelines/0091.pdf), which suggests that different definitions of ringleaders are used, or that the rule is not always enforced. Connor's database (used in the analysis) identifies the leader in each cartel, according to US DOJ reports, as a “cartelist mentioned in decision as a ringleader or a history of the case says one cartel member was the cartel disciplinarian/bully”.

²³ After Marvão's paper was widely circulated, a later paper uses the same EU specification on a shorter dataset and finds the same results (Hoang et al., 2014).

convicted between 1990 and 2008. They were often the largest cartel member(s) and formed agreements in markets with weak or no trade associations. The authors concluded that, although ringleaders were penalized more heavily after the introduction of the leniency policy, ringleader discrimination present in the 1996 EU leniency policy and removed from the 2002 version has not prevented the emergence of ringleaders.

On a related issue, Marvão and Spagnolo (2016) consider the debate on whether the introduction of criminal penalties in the EU, for individuals who engage in cartel activity, can strengthen antitrust enforcement. The authors document a recent phenomenon termed “leniency inflation” at the EU level and the fact that CEOs of convicted cartel members do not seem to be punished by shareholders as they often remain in the position or receive large severance packages. In addition, an empirical analysis of the criminal sanctions imposed in the USA suggests that repeat offenders are less likely to receive a prison sentence, in line with previous results suggesting that recidivist firms can use leniency to their own advantage. The authors interpret these results as suggestive of the need to introduce criminalization, in particular for infringements in the financial industry, possibly complemented by a moderate use of more expensive but proactive enforcement tools, such as screens and whistleblower rewards.

4 EXPERIMENTAL EVIDENCE ON LENIENCY

As previously discussed, cartels, like other white-collar crimes, are typically not observed unless they have been detected. Since every instance of collusion cannot be observed, interpreting an increase in the number of convicted cartels following a policy innovation as a “success” – an interpretation adopted by some in relation to the reform of the US leniency policy in 1993²⁴ – is an elementary logical mistake. An increase in the number of convictions may be generated by an increase in cartel formation, itself the result of more lenient law enforcement. As noted in Section 3.2, complex empirical methods need to be employed to try to understand whether the increased number of convicted cartels is associated with a fall or an increase in the total amount of such crimes in society. Not only are such studies necessarily complex and indirect (since the population of cartels is not directly observable, as is the case, for example, for violent crimes, most of which are reported), they are also of somewhat limited value when attempting to evaluate the effects of different policy designs that have not yet been implemented. Laboratory experiments are thus a crucial complementary empirical method because they overcome these drawbacks. They allow behavior to be observed in a controlled environment, including changes in the rate of overall cartel formation, and different policy designs to be tested at a reasonable cost.

Obviously, laboratory experiments themselves have several well-known drawbacks that offset their advantages to some extent. The results of laboratory experiments must therefore be carefully examined, particularly when assessing firm behavior based on the behavior of subjects in the laboratory. Because subjects are typically students and interaction is artificially simulated, the external validity of the results achieved cannot be taken for granted. With this caveat in mind, the following section reviews the available evidence from laboratory experiments on leniency and whistleblowers in competition law.

²⁴ See DOJ “Corporate Leniency Policy” (1993) at <http://www.justice.gov/atr/public/guidelines/0091.htm>.

4.1 Leniency, Rewards, Cartels and Prices: Early Studies

Apesteguia, Dufwenberg and Selten (2007) carried out the first laboratory experiment on leniency policies that we are aware of. They studied competitive outcomes in a one-shot homogeneous good Bertrand oligopoly with three firms and a discrete demand function. They embedded this market game in four legal frameworks: Ideal, Standard, Leniency and Bonus. In the “Ideal” framework, there was no antitrust law and communication across competitors (forming cartels) was not possible. In “Standard”, convicted firms faced fines equivalent to 10 percent of their revenue (accordingly, no fines were imposed if the firm had no revenue). In “Leniency”, firms that reported their participation in a cartel received a fine reduction (if they had some revenue and therefore faced a positive fine). And in “Bonus”, reporting cartel members received part of the fines paid by other firms as a reward. In this set up (homogeneous Bertrand and fines set at 10 percent of revenue), if a cartel member defected, its partners had zero revenue and therefore faced zero fines. For this reason, the presence or absence of leniency made no difference in terms of incentives to report, and strategically equivalent collusive sub-game perfect equilibria existed (in fact, full folk theorems hold) both in “Standard” and “Leniency”, sustained by the threat of reporting if a defection occurs.

The experimental analysis that tested the effects from the theoretical model confirmed that agents understand and use the threat of reporting to sustain collusion, more so in “Standard” than in “Leniency”, where both market prices and the percentage of cartel formation were lower. Additionally, “Leniency” was the framework that minimized the share of cartel formation. The analysis also did not find that deterrence increases with the introduction of rewards, since the “Bonus” framework presented the highest levels of market prices and cartel formation. However, in “Bonus”, incentives to report were stronger and there were no collusive equilibria sustained by the threat of reporting. This may suggest that the counterintuitive finding may not hold if subjects are allowed to gain experience. This leaves some room for follow-up work.

The stylized framework and particular setup used in this pioneering study raises some issues for the interpretation of its results. The oligopoly game in the experiment allowed for only one round of decisions, leaving agents no opportunity to learn the game. Coupled with the subtlety of the differences between “Standard”, “Leniency” and “Bonus”, it is possible that some of the counterintuitive results, such as agents not reacting to rewards, were driven by subjects not fully grasping the situation.

While Apesteguia et al.’s study tests the empirical relevance of theory, Hinlopen and Soeteven (2008b) approach the same issue but with a different methodology so as to make the lab look like the real world and thus, derive insights by analogy. They repeated the underlying oligopoly game and controlled for communication, allowing it to include different degrees of a range of electronically accepted market prices. Subjects were also free to choose whether or not to agree on a collusive price. When leniency was introduced, cartel members could only report and obtain a fine discount before an investigation was initiated. The first reporting party received full immunity and the second a 50 percent fine reduction; the remainder received no fine reduction at all. In this way, the study addressed both direct general deterrence and desistance effects. The study used the oligopoly model from Apesteguia et al.’s study as a stage game of a repeated game with an uncertain horizon, and added a small fixed cost of reporting to the legal framework. This cost had to be paid even when revenue was zero because a cartel partner undercut the price and took all customers. Although an additional

fixed cost/fine, limited to no-lenency treatments, would have further approximated real-world conditions, this positive reporting cost partly captured the real-world feature that, absent a lenency policy, a cheated-upon cartel member that reports is still subject to a fine. In this more realistic framework, the study confirmed the potential of the positive *ex ante* deterrence effects of the US lenency policy, restricted to the first “spontaneously” reporting party.

Contrary to what the first models of lenency assumed (e.g., Motta and Polo, 2003), Hinloopen and Soeteven showed that substantial cartel deterrence can be achieved with the introduction of a lenency policy that is only available to spontaneous reports before an investigation is opened. The average price in “Lenency” is significantly lower because cartels that do form are less successful in charging prices above the Nash equilibrium and because of the lower rate of decisions in favor of price discussions. This leads to a higher rate of defection and of price undercutting. Therefore, significantly fewer cartels are established and the life-span of cartels that were not deterred is reduced.

A second notable result of the study is that there exists a constant high rate of recidivism – the same percentage of detected and convicted cartels start colluding again, after some time, with or without lenency policies. Desistence (that is, specific deterrence) is not effective. The lack of desistance effects implied by recidivism may be a consequence of the absence of higher fines or the higher probability of detection for repeat offenders. Therefore, after a conviction, collusion remains practically as attractive as before for the convicted cartel. Unfortunately, the study did not consider rewards.

The above-mentioned studies focus on non-exploitable lenency policies.²⁵ In another study by Hinloopen and Soeteven (2008a), they used a similar setting (repeated Bertrand game, where subjects report before an investigation) but restricted it to a duopoly where communication was done through colored cards. They introduced an “exploitable” (overly generous) lenency policy treatment where agents could self-report and receive immunity from fines if they were the only reporter and a 90 percent reduction where they both reported. There were no penalties in the “benchmark” treatment, while in “antitrust”, cartels were detected with a 40 percent probability (much higher than the 15 percent in their other study) and they paid a fine equal to the cartel gains (compared to 10 percent of the revenue over the same period as in their other study). This simpler setting allowed Hinloopen and Soeteven to isolate the effects of exploitable lenency and non-exploitable lenency policy treatments.

The results in the paper showed that when there is an exploitable lenency policy, it is in fact exploited: 70 percent of the pairs reported simultaneously and there was some evidence that overt collusion became more appealing. It was also shown that a non-exploitable lenency policy treatment leads firms to turn to tacit collusion, which is not illegal and is thus free from fines. The non-exploitable lenency policy treatment led to an increase in overt collusion but of a much smaller magnitude than the exploitable lenency policy. The non-exploitable lenency policy treatment’s earnings were larger than in the benchmark treatment and no lower than in the exploitable lenency policy treatment. In conclusion, in this experiment, lenency policies always reduce welfare.

Hamaguchi, Kawagoe and Shibata (2009) considered the effects of cartel size (in terms of the number of members), the fine schedule and the degree of lenency (partial reduction, immunity or rewards) on the likelihood that a cartel is reported. In this study, subjects did not play a market game and did not choose prices or quantities. All subjects were initially assumed

²⁵ On exploitable lenency policies, see Spagnolo (2004).

or forced to be part of a cartel, but were given incentives to maintain collusion. The players were then left with the choice of whether to report collusion or not under different treatments, in which the leniency program is not necessarily strong enough to dissolve cartels. It was further assumed that cartels that are reported do not form again. The study found that the initial cartel was reported more frequently when the number of members was higher and that the frequency of reporting was neither affected by the fine schedule nor by whether only the first party or all parties that self-report were eligible for leniency. The study also found that the possibility of reporters receiving a reward had a large positive impact on dissolving cartel activity.

While these results on the likelihood of reporting are in themselves interesting, their interpretation in terms of the effects of leniency policies and their possible policy prescriptions is somewhat problematic. What matters for welfare is deterrence and prices, not the number of reports, which by themselves increase the workload of competition authorities and prosecution costs.²⁶ Experiments that include a market game show that there is a strong interdependence between the legal environment and the way firms behave in the market (Hinloopen and Soeteven, 2008b). This interaction is excluded by construction in the study by Hamaguchi et al. Therefore, it is not known if these reporting patterns would change if subjects were also involved in a market game as in reality. Also, ex ante deterrence effects and prices cannot be studied in this experiment because there is no cartel formation stage and no pricing decisions before or after the reporting stage.

4.2 Deviations, Pre-emption and the Level of Fines: Reaching the First Best

In Hinloopen and Soeteven's first study, subjects could report only in a simultaneous stage that took place after price choices were made public. Given this setup of the study, it was not possible for a cartel member that decides to abandon the cartel to "rush to the courtroom" before other cartel members realize they intend to do so. And it was therefore not possible to stop colluding and self-report before an opponent realizes that one of the cartel members wants to stop colluding and self-report. Yet this is a crucial feature of real-world leniency policies, both according to practitioners (e.g., Hammond, 2004a, 2004b, 2008) and according to theory: the "protection from fines" effect (Spagnolo, 2000a) and the "race to the courtroom effect" (Harrington, 2008) are severely limited by the impossibility of deviating from the cartel's price and reporting before the opponent realizes that deviation took place. Moreover, most leniency policies require the cessation of collusive conduct when applying for leniency, while the leniency application is kept secret (unless another firm applies) for quite some time so as to allow the competition authority to prepare for dawn raids and other actions. This means that leniency policies *require* secret deviation when secretly applying for leniency – something Hinloopen and Soeteven excluded. Finally, the fact that applications for leniency can only be submitted after the prices set by all competitors become public information makes the possibility of using leniency to punish price deviations particularly salient. As some have theorized, this may unduly enhance cartel stability (see Spagnolo, 2000a and Buccirossi and Spagnolo, 2006).

²⁶ We recognize that self-reporting increases cartel convictions, which is particularly advantageous in the case of a limited number of investigating officers. However, the ultimate focus of competition authorities should be improving welfare by increasing cartel deterrence and lowering prices.

To overcome these problems, which make it difficult to relate Hinloopen and Soeteven's results to real-world leniency policies, Bigoni et al. (2012) developed a dynamic experimental setting in which parties could apply for leniency, either before or after the price choices were observed by all players, in each stage game. This timing allowed a subject that wants to leave the cartel to both stop colluding on prices and apply for leniency confidentially before the other cartel members realize, as is possible in reality. This timing captured the "race to the courtroom" and "protection from punishment" effects (if you deviate on the price, you can apply for leniency at the same time so your competitors cannot punish your deviation by applying for leniency after they observe it). It also made it possible to disentangle and quantify reports linked to defections and reports linked to punishments. The setup also adopted a rematching methodology developed in the literature on experimental repeated games that allows subjects to face a constant discount factor and, most crucially, to play several supergames and learn.²⁷ It simplified the framework by using fixed fines so as to be able to control subjects' expectations on their level and how these change across treatments. The impact of these expectations on the effectiveness of leniency policies could therefore be studied.²⁸ Bigoni et al. also used a differentiated price game to avoid the non-generic and unrealistic discontinuities of the homogeneous-good Bertrand game (where a deviation implies zero profits – and in previous experiments zero fines – for all other firms), and a duopoly to minimize the risk highlighted by Holt (1995) that, with more than two subjects, punishment of deviators – which is crucial in studies of collusion – is biased or softened by the concern that the other, innocent subject will also be harmed by the punishment.

Bigoni et al. (2012) used this setup to study how standard antitrust enforcement (without leniency), leniency policies and monetary rewards for the first reporting party affect cartel formation and prices. They found that antitrust enforcement without leniency reduces cartel formation but increases cartel prices: subjects use costly fines as punishment against deviators. Leniency improves antitrust enforcement by strengthening deterrence, as fewer cartels are formed and existing cartels that are detected through leniency do not form again (leniency eliminates recidivism²⁹). However, leniency policies also stabilize surviving cartels: subjects appear to anticipate the lower post-conviction prices and lack of recidivism after self-reports or leniency. Therefore, overall average prices do not fall significantly. Conversely, with rewards, prices rapidly fall to the competitive level. Overall, the results suggest a strong cartel deterrence potential for well-run leniency policies, where firms self-report before an investigation is opened. The results also suggest that rewards should be introduced to obtain substantial welfare gains in terms of lower prices.

In a subsequent study, Bigoni et al. (2015) used this same setup to study the effect of separately changing the level of the fines and the probability of exogenous detection on cartel deterrence, with and without leniency. For occasional crimes committed by single and risk-neutral subjects, changing the mix between fines and exogenous probability of detection, keeping the expected fine constant, should not affect deterrence. The paper developed a model

²⁷ See, e.g., Dal Bó (2005), Blonski et al. (2011), Dal Bó and Fréchette (2011) and Bigoni et al. (2015).

²⁸ When fines are set as a share of the profits realized in a previous period, as in Hinloopen and Soeteven (2008b), it is hard for subjects to predict what the fine will be and for the experimenter to control for what subjects' expectations are, because cartels are often detected and fined after they have stopped sustaining high prices. The fine is often therefore a fraction of competitive, rather than collusive, profits. This feature makes it impossible to control for the level of fines and study how this interacts with the leniency policy.

²⁹ Against Hinloopen and Soeteven (2008b).

showing that in a dynamic multi-agent setup, this equivalence is lost and fines are much more important with leniency. The experiment confirmed the theoretical finding. Without leniency, the probability of exogenous detection and fines both have similar effects for deterrence. With leniency policies in place, the absolute level of the fine is much more important in producing deterrence, while the probability of exogenous detection becomes practically irrelevant. This indicates that deterrence is mainly driven by “distrust” or strategic risk, that is, by the fear of partners deviating and reporting. This study even found a large deterrence effect of fines in the presence of a leniency policy when the probability of exogenous detection is zero. As theorized by Spagnolo (2004), this implies that the “distrust” deterrence channel is powerful and that the first-best scenario (full deterrence with zero deadweight/inspection costs) could now be achieved at finite levels of fines. It also implies that recently voiced concerns that the large number of leniency applications may be reducing antitrust effectiveness by exhausting the resources of competition authorities, making it impossible for them to undertake random industry audits, may be misplaced (see Abrantes-Metz, 2013). On the contrary, these findings suggest that the efficiency of competition law enforcement can be considerably improved by strengthening sanctions and the management of the leniency policy while reducing the expenditure of competition authorities’ resources on random inspections.

Of course, it is important to ensure that these results are robust before translating them into policy prescriptions. Positive news in this respect is found in a recent experiment by Chowdhury and Wandschneider (2013). This study also considered the effect of changing the mix between fines and the exogenous detection probability in the absence and presence of a leniency policy, as studied by Bigoni et al. (2015), although it did not consider the case of zero probability of detection. However, it did so in an environment similar to the one in Hinloopen and Soetevent’s (2008b) study, where matching was fixed and cartels could only be reported after price choices were made public, so that – as in Hinloopen and Soetevent’s work – the “protection from punishment” and “race to the courtroom” effects could hardly be active. Bigoni et al.’s finding was confirmed by this experiment: increasing the absolute fine and reducing the probability of exogenous detection (absent self-reporting) increased the deterrent effect of leniency policies in this environment also. The conclusion that the efficiency of competition law enforcement can be improved by strengthening sanctions and the management of the leniency policy while reducing competition authorities’ efforts in conducting random inspections of industries seems rather robust.

Additional support for Spagnolo’s (2004) and Bigoni et al.’s (2015) “strategic risk/distrust” deterrence channel is found in Kindsgaab (2015), where an experiment that extends Bigoni et al. (2015) is run with the purpose to examine how changes in the expected fines and reporting costs affect that type of deterrence. The author finds that suboptimal leniency policies increase deterrence more than a fine policy (without leniency) with an equivalent expected fine, and that this result is exclusively driven by this novel distrust channel.

4.3 Additional Issues

4.3.1 Ringleaders

One debated issue is whether ringleaders should be excluded from leniency policies (as in the USA) or included (as in the EU). On the one hand, excluding ringleaders from the leniency policy may increase deterrence by introducing “free riding” on who should lead. Excluding

ringleaders may discourage firms from taking the lead and induce them to wait for others to do so, thereby delaying and reducing cartel formation. On the other hand, this policy may reduce deterrence by creating one firm that can be trusted by the others as it will never (be able to) “run to the courtroom” and report them.

Bigoni et al. (2012) undertook a preliminary investigation of this trade-off by introducing treatments where the ringleader, defined as the subject that first asked the others to communicate, could not apply for leniency. This was announced to subjects, who therefore knew that they would lose the opportunity to receive leniency if they communicated first. The authors found that in treatments where the initiator of the cartel could not apply for leniency, the deterrence effect of leniency is unaffected, although prices increase. They argued, however, that this was a preliminary result that should be treated with caution, as the experimental setup was not explicitly designed to address this question and was particularly unfavorable to excluding ringleaders. With a duopoly, excluding the ringleader leaves only one party able to report and obtain leniency, which eliminates the fear of others reporting, which is, according to Spagnolo (2004), a crucial determinant of deterrence. Bigoni et al. therefore invited more work on the subject. The invitation was taken up by a number of authors.

Hesch (2012) used a simplified version of Hinloopen and Soetevent’s (2008b) model where reporting could only take place after price choices became public and where liability expired after each period. He introduced a ringleader role, which was assigned randomly by a computer in each period. It was found that, in treatments where the randomly assigned ringleader was not allowed to apply for leniency, cartel formation was more intense and prices were higher. Unfortunately, an exogenous and random assignment of the role of ringleader eliminates, by design, coordination problems in the formation of the cartel, which is where a positive effect of excluding ringleaders could occur. By removing the possibility that coordination issues could be worsened by the exclusion of ringleaders, inducing subjects to delay or avoid taking the lead hoping that others would do it first, the experiment allowed for only the negative effects of the policy. This reduces the validity of the result.

Wandschneider (2014) improved the mechanism to identify the leader. In his setup, the ringleader was the subject whose suggested cartel price during the communication stage had been accepted by the two other group members. As in earlier work (Bigoni et al., 2012), this made the identity of the (at least partial) leader endogenous. A form of “free-riding effect” could then in principle present itself, not in the form of delayed or reduced cartel formation but in the form of lower prices suggested by those who do not want to be the leader, which could possibly induce lower cartel prices.

The study found that more cartels are formed when the leader is not able to obtain leniency. However, it also found that prices do not increase when ringleaders are excluded from the leniency policy, which might be due to the above-mentioned free-riding effect. An in-depth analysis of behavior in the price proposal stage is needed to verify this conjecture. Finally, it found that ringleader exclusion destabilizes the collusive agreement, as more firms deviate. This was expected, as this study follows Hinloopen and Soetevent (2008b) in only allowing applications for leniency after price defections are made public. As we have explained, this ensures that the leniency policy is mainly used to discipline price defections, as it excludes the pro-competitive effects linked to the optimal “deviate and report” strategies. Since excluding ringleaders allows only the deviator and one more firm to report, the punishment for non-ringleaders that deviate on price is reduced; when they report, they expect half of the fine reduction instead of one-third. Before drawing any policy conclusions from these results, it is

therefore important to wait for more realistic studies that allow subjects to apply for leniency when deviating on prices and be rematched to play several supergames and learn.

More recently, Clemens and Rau (2014) studied the ringleader issue in a reduced form participation-revelation game in which ringleaders may or may not emerge. They implemented a cartel formation game where the cartel is established in a multi-stage decision game preceded by a communication stage. If some cartel members chose to open a communication window that was not necessary for the cartel to be formed, these cartel members became the ringleaders. The experimental design did not include any form of market interaction, whether static or dynamic, nor pricing decisions. Subjects that chose to take part in a cartel were then always bound to the joint profit-maximizing strategy, while outside firms played best response. They then implemented treatments without leniency, with leniency open to all, and with leniency only open to non-ringleaders. They found that excluding ringleaders from obtaining leniency reduced the number of reports, increased the number of cartels formed, and even increased the number of subjects becoming ringleaders. They concluded that excluding ringleaders from the leniency policy is likely to reduce its effectiveness.

These results (complementing the empirical work of Davies and De, 2013 and the theoretical work on ringleaders mentioned in Section 2.3) are instructive, as they isolate the effect of ringleader exclusion on reporting, from their interaction with market strategies. However, in terms of evaluating the effectiveness of a leniency policy, they suffer from a similar limitation as the study by Hamaguchi et al. (2009). As previously discussed in relation to that paper, it is difficult to interpret and translate into policy prescriptions the results of an experimental design that does not include any form of market interaction. Clemens and Rau (2014) took the view that not including a market game was “necessary as defection from the cartel price by a shirking firm might influence the decision to form a cartel as much as the possibility to opt for leniency”. Indeed, we know from the previously described experiments that market behavior and reporting behavior interact in important ways. From a policy point of view, however, we are interested in these interactions, as it is cartel formation and prices that determine changes in welfare, not the number of reports (which in themselves typically lower welfare by increasing prosecution costs). If we exclude market interactions from the design, it becomes difficult to understand if and how the measured reporting behavior would change in the presence of market interactions, and how market outcomes and welfare are likely to be affected by leniency.

To conclude, taken together, these available experimental results suggest that ringleaders should be allowed to apply for and obtain leniency. However, given the caveats in all these studies, further research appears necessary to investigate the robustness of this conclusion.

4.3.2 Leniency and auctions

Hamaguchi et al. (2007) studied collusion in a repeated procurement auction game and the effectiveness of leniency policies in that environment. They considered cartel creation at first-price sealed-bid auctions and allowed for unrestricted communication before bidding. The experiment allowed for five competitors and the formation of partial cartels. In addition, the competition authority could detect individual cartel members (but not the entire cartel) and the fine imposed was a share of the individual’s gross earnings in the last three periods. No communication was allowed before the bid in the “benchmark” treatment, whereas in “communication”, a three-minute chat where subjects decided whether or not to enter the chatroom preceded the possible bid. In “antitrust”, communication was allowed and there

was a 15 percent probability of detection by a competition authority. In “communication”, virtually all bids were set at the monopoly price, so bidders clearly colluded and did not cheat on the agreement reached in this phase. Leniency policies turned out to be ineffective in decreasing the number of cartels in the auctions, and the average winning bid did not change. However, there was some evidence that leniency policies may be effective to dissolve pre-existing collusion and decrease the contract price. In “antitrust”, most of the pre-collusive groups bid their reserve price and were then dissolved by defectors before the end of the experiment.

Hinloopen and Onderstal (2013) studied cartel formation and leniency policies at first-price sealed-bid and English auctions. In their experiment, each subject started by choosing between “yes” or “no” buttons that indicated their willingness to join a possible cartel. They were then told whether a cartel formed, but not about individual votes. If a cartel was established, a winner was randomly assigned by the computer and was the only subject who could submit a bid. The highest bidder won the object. In subsequent rounds, subjects needed to bid higher than the winning bid, and the rounds ended when no subject bids or when one bids the maximum possible bid. There was no competition authority in “agreement”, but the “detect and punish” and “leniency” treatments entailed a 15 percent chance of detection and prosecution. In the latter, firms could also report the cartel once the auction ended for a small cost, and they did so ignorant of the other player’s reporting decision. Hamaguchi et al.’s (2007) result on the ineffectiveness of the leniency policy in first-price sealed-bid auctions was corroborated. Nonetheless, in English auctions, a traditional antitrust policy (with no leniency policy) seems able to deter and destabilize cartels, but it also has the negative effect of reducing the average winning bid (that is, the price). Although the introduction of a leniency policy seems to have had no impact on cartel formation or recidivism, it did have two undesirable effects: it increased cartel stability and reduced the winning cartel bid, in line with the results from Bigoni et al. (2012).

4.3.3 Leniency after an investigation is initiated and avoidance activities

All the experimental work discussed to this point in the chapter does not specify whether or not an investigation of the cartel had been started at the time a leniency application is made. The assumed positive probability of exogenous detection can be interpreted both as the probability of a successful investigation and as the probability that an existing investigation, started with the formation of the cartel, will be successful. The presence of robust deterrence effects in many of these experiments demonstrates that the assumption on which early studies of leniency policies are based – that programs restricted only to spontaneous reports before an investigation is open cannot be effective – is incorrect both logically and empirically. These experiments cannot tell us, however, how opening leniency policies to reports coming after an investigation is opened or announced will affect deterrence and welfare.

This question was the focus of a study by Dijkstra, Haan and Schoonbeek (2014), in which firms could apply for leniency once an antitrust investigation had been announced and could also communicate freely. In the common setting of a repeated and homogeneous Bertrand duopoly, if firms chose to communicate and set prices, an investigation may (or not) be opened. Subjects could apply for leniency once they learned about this, thereby ensuring conviction. Otherwise, conviction occurred with some probability. If convicted, a fixed fine was paid. The experiment showed that individuals are able to fix and keep prices high by agreeing on prices and reporting and by agreeing on future communication strategies. Some

evidence of desistance and destabilization effects, due to the leniency policy, was found in the very short term, but these disappeared over time.

Finally, an interesting recent experiment by Chowdhury and Wandschneider (2013) looked at the effect of leniency policies when firms can invest in costly avoidance activities, an important and under-researched topic in competition law. They augmented the stage game from their earlier work with the possibility, in some treatments, of cartel members undertaking a costly investment that would permanently reduce the (absolute) fine they would face in future periods if convicted. The authors found that avoidance activities increase cartel formation (by risk-averse subjects) and that firms that invest in avoidance charge higher prices. They also found that such firms deviate and self-report more often when a leniency policy exists. This indicates that in the presence of a leniency policy, some firms use avoidance to reduce their punishment for price deviations. This is what should be expected in a setup similar to that of Hinloopen and Soeteven (2008b), where firms can only self-report after prices (deviations) become known. In such a setup, the leniency policy mostly acts as a punishment device. Understanding how their results would change in a more realistic setup, where firms can also report before their price deviations become common knowledge, appears to be an exciting avenue for further research.

5 CONCLUSION

There is no doubt about the increasing importance of leniency policies for competition authorities' daily enforcement work. This is clearly reflected in the growing number of firms applying for leniency reductions in exchange for information and cooperation with an ongoing investigation.

Having reviewed the literature on leniency programs, it seems fair to conclude that if these are well-designed and well-administered programs, they can be a powerful policy instrument to combat illegal cartels and should be present in the toolkit of any competition authority, regardless of their budget.

However, the literature has also pointed out that if these information revelation mechanisms are poorly designed and/or too generously administered, they can have serious counter-productive effects, by providing an easy way for cartelists to escape or reduce fines, and potentially encouraging cartels that would not otherwise form, while merely making it easier for competition authorities to detect and prosecute cartels. In addition, these programs seem to be used by cartel members as punishment strategies (grim-trigger versus "stick and carrot") to sustain collusion.

There are clear signs that the perceived "pros" of leniency are leading the European Commission to overuse leniency, as if it was a form of plea bargaining. This is natural since the number of convicted cartels is used as a performance measure, and may be efficient in some specific (but rare) cases given that plea bargaining is not available and settlements can only award a limited discount on the fine. The fact that leniency reductions have been granted in 52 percent of all EC cartel fines (1998–2014), and that this percentage, corresponding to an average of four leniency recipients per cartel, is on the rise, reveals that this bias is increasing, together with the excessive deadweight loss for society.

Evaluating how these policies are implemented in reality, and how their design and management could be improved, is therefore crucial. A much stricter implementation of

leniency policies, complemented by strengthened sanctions (including damage payments) and possibly a moderate use of more expensive but proactive enforcement tools, such as screens, appears to be the way forward.

A large and growing body of research has studied the impact of leniency programs on cartel stability, most of which agreeing on some of the optimal features of leniency programs. The theoretical literature suggests that these features include: full amnesty and limited liability for the first reporting firm only, limited leniency reductions to subsequent reporters if the authorities don't have sufficient information (or sufficient budget) to guarantee a conviction, enabling the cartel leader to be awarded leniency but increasing the punishment of repeat offenders, transparency and predictability in the setting of leniency reductions, and possibly a reward scheme to (individual) whistleblowers, financed by the fines imposed on all other cartel members. Implementing such policies will avoid using leniency programs as a disciplining device to hinder defections, particularly in the context of multi-market contact, or as a strategic device to escape fines and damage the reporter's competitors, which is mostly important when there is a large asymmetry of information between firms and the competition authority (causing them to commit type I and type II errors) or amongst cartel members.

Recent theories also suggest that the opening of an investigation may exhibit a knock-on effect on leniency applications across multiple markets; and call for cooperation amongst competition authorities in such a context.

The existing empirical studies provide mixed results. Our conclusion from reviewing the empirical work is that much more empirical work is required. Judging from the very limited empirical evidence available, it is still not well established whether leniency policies, as currently designed and implemented in different countries, are doing any more than facilitating competition authorities' work. That is, it is unclear whether they are actually increasing welfare by generating a strong deterrence effect, or whether they are actually reducing welfare through the larger administration and prosecution costs they generate, without any compensating increase in deterrence. The most favorable evidence available is for the United States, where sanctions are much tougher, and this is consistent with what theory would predict. But overall, the evidence is in general rather weak.

An increasing number of experimental studies clearly demonstrate that the assumption on which some early economic analyses were based – that leniency policies are only effective if they allow reports from firms under investigation – is not only ad hoc and unjustified, but also empirically counterfactual. Although this is not to say that, given constraints on sanctions and rewards, it is not optimal to open leniency policies to reports after an investigation. The bulk of experiments also suggests, consistent with the available empirical evidence, that cartel deterrence effects of well-designed and well-administered leniency policies tend to be positive – whether or not the policy is open to reports after an investigation opened – but rather modest unless sanctions for non-applicants are severe or monetary rewards are introduced. Most recent experiments suggest that severe sanctions are the crucial precondition for the effectiveness of a leniency policy, allowing it to produce substantial cartel deterrence effects even when the probability of a cartel being detected without reports is zero.

Experiments also show that subjects quickly understand how to play these schemes, if they can be played, so that poorly designed and loosely administered real-world leniency policies are likely to reduce social welfare considerably.

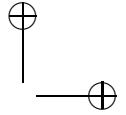
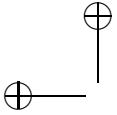
Some experiments tend to have rather loose connections with both the theory and the practice of leniency policies, making it hard to use their results as guidance for policy-making.

Future experimental work should pay more attention to both theory and reality. Several open questions are waiting for more careful examination, starting with the introduction of fines or damage payments that are a function of accumulated cartel profits.

The lack of stronger evidence – whether in favor or against the hypothesis that leniency policies are increasing cartel deterrence and with it social welfare – is undoubtedly linked to the difficulty of identifying how the total population of cartels changes when leniency policies are introduced or modified. But it is also clearly linked to an endemic lack of data. The development of meaningful research on leniency would be facilitated if competition authorities or agencies in charge of supervising them start to implement more consistent data collection and data disclosure policies.

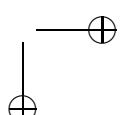
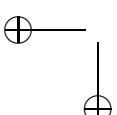
REFERENCES

- Abrantes-Metz, R. (2013), "Proactive vs reactive anti-cartel policy: The role of empirical screens", *Working Paper*.
- Angelucci, C. and M. Han (2011), "Private and public control of management", *Amsterdam Center for Law & Economics Working Paper*, No. 2010-14.
- Apesteguia, J., M. Dufwenberg and R. Selten (2007), "Blowing the whistle", *Economic Theory*, **31**, 143–66.
- Arlman, S. (2005), "Crime but no punishment: An empirical study of the EU's 1996 leniency notice and cartel fines in Article 81 proceedings", MSc thesis, University of Amsterdam.
- Armantier, O. and A. Boly (2013), "Comparing corruption in the laboratory and in the field in Burkina Faso and in Canada", *The Economic Journal*, **123**(573), 1168–87.
- Asker, J. (2010), "Leniency and post-cartel market conduct: Preliminary evidence from parcel tanker shipping", *International Journal of Industrial Organization*, **28**, 407–14.
- Aubert, C., W. Kovacic and P. Rey (2006), "The impact of leniency programs on cartels", *International Journal of Industrial Organization*, **24**(6), 1241–66.
- Bageri, V., Y. Katsoulacos and G. Spagnolo (2013), "The distortive effects of antitrust fines based on revenue", *The Economic Journal*, **123**(572), 545–57.
- Becker, G. (1968), "Crime and punishment: An economic approach", *Journal of Political Economy*, **76**(2), 169–217.
- Bigoni, M., M. Casari, A. Skrzypacz and G. Spagnolo (2015), "Time horizon and cooperation in continuous time", *Econometrica*, **83**(2), 587–616.
- Bigoni, M., S.-O. Fridolfsson, C. Le Coq and G. Spagnolo (2012), "Fines, leniency, and rewards in antitrust", *RAND Journal of Economics*, **43**(2), 368–90.
- Bigoni, M., S.-O. Fridolfsson, C. Le Coq and G. Spagnolo (2015), "Trust, leniency and deterrence", *Journal of Law, Economics and Organization*, **31**(4), 663–89.
- Blatter, M., W. Emond and S. Sticher (2014), "Optimal leniency programs when firms have cumulative and asymmetric evidence", *Discussion Paper*, University of Bern.
- Blonski, M. and G. Spagnolo (2014), "Prisoners' other dilemma", *International Journal of Game Theory*, **44**(1), 61–81.
- Blonski, M., P. Ockenfels and G. Spagnolo (2011), "Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence", *American Economic Journal: Microeconomics*, **3**(3), 164–92.
- Bloom, M. (2007), "Despite its great success, the EC Leniency Programme faces great challenges", in C.-D. Ehlermann and I. Atanasiu (eds), *European Competition Law Annual 2006: Enforcement of Prohibition of Cartels*, Oxford: Hart Publishing.
- Bos, I. and J. Harrington (2010), "Endogenous cartel formation with heterogeneous firms", *The RAND Journal of Economics*, **41**(1), 92–117.
- Bos, I. and F. Wandschneider (2012), "Cartel ringleaders and the corporate leniency program", *CCP Working Paper*, No. 11–13.
- Bos, I., S. Davies, J. Harrington and P. Ormosi (2016), "Does enforcement deter cartels? A tale of two tails", *CCP Working Paper*, No. 14-6 v2.
- Breitmoser, Y. (2015), "Cooperation, but no reciprocity: Individual strategies in the repeated prisoner's dilemma", *American Economic Review*, **105**(9), 2882–910.
- Brenner, S. (2009), "An empirical study of the European corporate leniency program", *International Journal of Industrial Organization*, **27**, 639–45.
- Brenner, S. (2011), "Self-disclosure at international cartels", *Journal of International Business Studies*, **42**(2), 221–34.
- Brisset, K and L. Thomas (2004), "Leniency program: A new tool in competition policy to deter cartel activity in procurement auctions", *European Journal of Law and Economics*, **17**(1), 5–19.



88 *Handbook of game theory and industrial organization: applications*

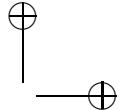
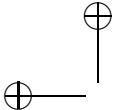
- Buccirossi, P and G. Spagnolo (2006), "Leniency programs and illegal transactions", *Journal of Public Economics*, **90**(6–7), 1281–97.
- Buccirossi, P., C. Marvão and G. Spagnolo (2015), "Leniency and damages", *CEPR Discussion Paper*, No. DP 10682.
- Carree, M., A. Günster and M.P. Schinkel (2010), "European antitrust policy 1957–2004: An analysis of Commission decisions", *Review of Industrial Organization*, **36**(2), 97–131.
- Chen, Z (2008), "Cartel organization and antitrust enforcement", *CCP Working Paper*, No. 08-21.
- Chen, Z and P. Rey (2013), "On the design of leniency programs", *The Journal of Law and Economics*, **56**(4), 917–57.
- Chen, Z., S. Ghosh and T. Ross (2015), "Denying leniency to cartel instigators: Costs and benefits", *International Journal of Industrial Organization*, **41**, 19–29.
- Choi, J.P. and H. Gerlach (2012), "Global cartels, leniency programs and international antitrust cooperation", *International Journal of Industrial Organization*, **30**(6), 528–40.
- Chowdhury, S. and F. Wandschneider (2013), "Anti-trust and the 'Beckerian proposition': The effects of investigation and fines on cartels", *CCP Working Paper*, No. 13-9.
- Clemens, G. and H. Rau (2014), "Do leniency policies facilitate collusion? Experimental evidence", *DICE Discussion Papers*, No. 130.
- Cloutier, M. (2011), "An empirical investigation of the U.S. corporate leniency programme", *Working Paper*, Queens University, Canada.
- Combe, E. and C. Monnier (2009), "Fines against hard core cartels in Europe: The myth of overenforcement", *Antitrust Bulletin*, **56**(2), 235–75.
- Connor, J. (2007), "Global antitrust prosecutions of international cartels: Focus on Asia", *World Competition*, **31**(4), 575–605.
- Connor, J. (2010), "Recidivism revealed: Private international cartels 1991–2009", *Competition Policy International*, **101** (Autumn).
- Connor, J. (2013), "Cartel fine severity and the European Commission: 2007–2011", *European Competition Law Review*, **34**, 58–77.
- Connor, J. and D. Miller (2013), "Determinants of U.S. antitrust fines of corporate participants of global cartels", *Working Paper*.
- Cooter, R. and T. Ulen (1988), *Law and Economics*, Glenview, IL: Scott, Foresman & Co.
- Dal Bó, P. (2005), "Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games", *American Economic Review*, **95**(5), 1591–604.
- Dal Bó, P. and G. Fréchette (2011), "The evolution of cooperation in infinitely repeated games: Experimental evidence", *American Economic Review*, **101**(1), 411–29.
- Dargaud, E. and A. Jacques (2016), "Endogenous firms' organization, internal audit and leniency programs", *Working Paper*, presented at the EARIE Conference 2016.
- Davies, S. and O. De (2013), "Ringleaders in larger number asymmetric cartels", *Economic Journal*, **123**(572), F524–F544.
- Davies, S., P. Ormosi and M. Graffenberger (2014), "Mergers after cartels: How markets react to cartel breakdown", *Working Paper*.
- De, O. (2010), "Analysis of cartel duration: Evidence from EC prosecuted cartels", *International Journal of the Economics of Business*, **17**(1), 33–65.
- Dijkstra, P., M. Haan and L. Schoonbeek (2014), "Leniency programs and the design of antitrust: Experimental evidence with rich communication", *Working Paper*.
- Dominte, O., D. Ţerban and A. Dima (2013), "Cartels in EU: Study on the effectiveness of leniency policy", *Management & Marketing*, **8**(3), 529–552.
- Dong, A., M. Massa and A. Žaldokas (2014), "Busted! Now what? Effects of cartel enforcement on firm policies", *INSEAD Working Paper*, No. 2014/38/FIN.
- Ellis, C and W. Wilson (2003), "Cartels, price-fixing, and corporate leniency policy: What doesn't kill us makes us stronger", mimeo.
- Feeß, E. and M. Walzl (2004), "Self-reporting in optimal law enforcement when there are criminal teams", *Economica* **71**(283), 333–48.
- Feeß, E. and M. Walzl (2010), "Evidence dependence of fine reductions in corporate leniency programs", *Journal of Institutional and Theoretical Economics*, **166**(4), 573–90.
- Gärtner, D (2013), "Corporate leniency in a dynamic world: The preemptive push of an uncertain future", *Working Paper*.
- Gärtner, D. and J. Zhou (2012), "Delays in leniency application: Is there really a race to the enforcer's door?" *GESY Discussion Paper*, No. 395.
- Gerlach, H. (2013), "Self-reporting, investigation and evidentiary standards", *Journal of Law and Economics*, **56**(4), 1061–90.
- Grossman, G. and M. Katz (1983) "Plea bargaining and social welfare", *American Economic Review*, **73**, 749–57.
- Hamaguchi, Y., T. Ishikawa and M. Ishimoto et al. (2007), "An experimental study of procurement auctions with leniency programs", *CPRC Discussion Paper Series*, No. CPDP-24-E.
- Hamaguchi, Y., T. Kawagoe and A.O. Shibata (2009), "Group size effects on cartel formation and the enforcement power of leniency programs", *International Journal of Industrial Organization*, **27**(2), 145–65.
- Hammond, S.



- (2004a), "Detecting and deterring cartel activity through an effective leniency program", presented at the International Workshop on Cartels, Brighton, 21–22 November.
- Hammond, S. (2004b), "Cornerstones of an effective leniency program", presented at the ICN Workshop on Leniency Programmes, Sydney, 22–32 November.
- Hammond, S. (2008), "Cornerstones of an effective cartel leniency programme", *Competition Law International*, 4(2).
- Harrington, J. (2008), "Optimal corporate leniency programs", *Journal of Industrial Economics*, 56(2), 215–46.
- Harrington, J. (2013), "Corporate leniency programs when firms have private information: The push of prosecution and the pull of pre-emption", *The Journal of Industrial Economics*, 61(1), 1–27.
- Harrington, J. and M.-H. Chang (2009), "Modelling the birth and death of cartels with an application to evaluating antitrust policy", *Journal of the European Economic Association*, 7, 1400–435.
- Harrington, J. and M.-H. Chang (2013), "Endogenous antitrust enforcement in the presence of a corporate leniency program", *Cleveland State University Working Paper*, No. 26/2012.
- Harrington, J. and M.-H. Chang (2015), "When should we expect a corporate leniency program to result in fewer cartels?", *The Journal of Law and Economics*, 58(2), 417–49.
- Harrington, J. and J. Chen (2006), "Cartel pricing dynamics with cost variability and endogenous buyer detection", *International Journal of Industrial Organization*, 24(6), 1185–212.
- Harsanyi, J. and R. Selten (1998), *A General Theory of Equilibrium Selection in Games* Cambridge MA: MIT Press.
- Herre, J., W. Mimra and A. Rasch (2012), "Excluding ringleaders from leniency programs", *Working Paper*.
- Hesch, M. (2012), "The effects of ringleader discrimination on cartel stability and deterrence – Experimental insights", *Journal of Advanced Research in Law and Economics*, 3(1), 26–39.
- Hinloopen, J. and S. Onderstal (2013), "Going once, going twice, reported!", *Tinbergen Institute Discussion Paper*, No. TI 2009-085/1.
- Hinloopen, J. and A. Soeteven (2008a), "From overt to tacit collusion: Experimental evidence on the adverse effects of corporate leniency programs", *Tinbergen Institute Discussion Paper*, No. TI 2008-059/1.
- Hinloopen, J. and A. Soeteven (2008b), "Laboratory evidence on the effectiveness of corporate leniency programs", *RAND Journal of Economics*, 39(2), 607–16.
- Hoang, C.G., K. Huschelrath, U. Laitenberger and F. Smuda (2014), "Determinants of self-reporting under the European corporate leniency program", *International Review of Law and Economics*, 40, 15–23.
- Holt, C. (1995), "Industrial organization: A survey of laboratory research", in J.H. Kagel and A.E. Roth (eds), *The Handbook of Experimental Economics*, Princeton, NJ: Princeton University Press.
- Houba, H., E. Motchenkova and Q. Wen (2011), "Antitrust enforcement and marginal deterrence", *Tinbergen Institute Discussion Paper*, No. 11/166/1.
- Houba, H., E. Motchenkova and Q. Wen (2015), "The effects of leniency on cartel pricing", *The B.E. Journal of Theoretical Economics*, 15(2), 351–89.
- Innes, R. (1999), "Self-policing and optimal law enforcement when violator remediation is valuable", *Journal of Political Economy*, 107(6), 1305–25.
- Kaplow, L. and S. Shavell (1994), "Optimal law enforcement with self-reporting of behavior", *Journal of Political Economy* 102(3), 583–606.
- Kindsgrab, P. (2015), "Deterrence and suboptimal leniency: An experiment", BSc thesis, University of Mannheim.
- Klein, G. (2011), "Cartel destabilization and leniency programs – Empirical evidence", *ZEW Discussion Paper*, No. 10-107.
- Kobayashi, B. (1992), "Deterrence with multiple defendants: An explanation for 'unfair' plea bargains", *RAND Journal of Economics*, 23(4), 507–17.
- Kofman, F. and J. Lawarrée (1996), "A prisoner's dilemma model of collusion deterrence", *Journal of Public Economics* 59(1), 117–36.
- Lefouili, Y. and C. Roux (2012), "Leniency programs for multimarket firms: The effect of amnesty plus on cartel formation", *International Journal of Industrial Organization*, 30(6), 624–40.
- Lewis, D (2006), Speech of David Lewis, Chairperson of the Competition Tribunal of South Africa on Competition and Development, 2 May 2006, Cape Town, South Africa.
- Malik, A (1993), "Self-reporting and the design of policies for regulating stochastic pollution", *Journal of Environmental Economics and Management*, 24(3), 241–57.
- Marvão, C. (2014), "Heterogeneous penalties and private information", *Konkurrensverket Series in Law and Economics Working Paper*, No. 2014:1.
- Marvão, C. (2015), "The EU Leniency Programme and recidivism", *Review of Industrial Organization*, 48(1), 1–27.
- Marvão, C. and G. Spagnolo (2016), "Should price fixers finally go to prison? Criminalization, leniency inflation and whistleblower rewards in the EU", *Working Paper*
- Marvão, C. and G. Spagnolo (in progress, September 2017), "Deterrence effects of leniency programs: another tale of tails"
- Marx, L. and J. Zhou (2014), "The dynamics of mergers among (ex)co-conspirators in the shadow of cartel enforcement", *TILEC Discussion Paper*, No. 2014-013.

90 *Handbook of game theory and industrial organization: applications*

- Marx, L., C. Mezzetti and R. Marshall (2015), "Antitrust leniency with multiproduct colluders", *American Economic Journal: Microeconomics*, **7**(3), 205–40.
- Miller, N. (2009), "Strategic leniency and cartel enforcement", *American Economic Review*, **99**(3), 750–68.
- Motchenkova, E. (2005), "Optimal enforcement of competition law", PhD thesis, Tilburg University.
- Motchenkova, E. and T. Ghebrihiwet (2010), "Leniency programs in the presence of judicial errors", *TILEC Discussion Paper*, No. 2010-030.
- Motchenkova, E. and D. Leliefeld (2010), "Adverse effects of corporate leniency programs in view of industry asymmetry", *Journal of Applied Economic Sciences*, **5**(2), 114–28.
- Motchenkova, E. and R. van der Laan (2011), "Strictness of leniency programs and asymmetric punishment effect", *International Review of Economics*, **58**(4), 401–31.
- Motta, M. and M. Polo (2003), "Leniency programs and cartel prosecution", *International Journal of Industrial Organization*, **21**(3), 347–79.
- Pavlova, N. and A. Shastikov (2014), "Effects of hostility tradition in antitrust: Leniency programs and cooperation agreements", *Basic Research Program Working Paper Series*, No. BRP 58/EC/2014.
- Pinna, A. (2010), "Optimal leniency programs in antitrust", Università di Cagliari and Università di Sassari, *CRENoS Working Paper*, No. 2010/18.
- Posner, R. (1976), *Antitrust law: An Economic Perspective*, Chicago, IL: University of Chicago Press.
- Reinganum, J. (1988), "Plea bargaining and prosecutorial discretion", *The American Economic Review*, **78**(4), 713–28.
- Rey, P. (2003), "Towards a theory of competition policy", in M. Dewatripont, L.P. Hansen and S.J. Turnovsky (eds), *Advances in Economics and Econometrics: Theory and Applications*, Cambridge, UK: Cambridge University Press.
- Sauvagnat, J. (2014), "Are leniency programs too generous?", *Economics Letters*, **123**(3), 323–6.
- Sauvagnat, J. (2015), "Prosecution and leniency programs: The role of bluffing in opening investigations", *The Journal of Industrial Economics*, **63**(2), 313–38.
- Shastikov, A. and S. Avdasheva (2011), "Introduction of leniency programs for cartel participants: The Russian case", *Antitrust Chronicle*, **8**(2).
- Silbye, F. (2010), "Asymmetric evidence and optimal leniency programs", PhD thesis, University of Copenhagen, 41–61.
- Spagnolo, G (2000a), "Optimal leniency programs", *F.E.E.M. Nota di Lavoro No. 42.00*, Fondazione ENI "Enrico Mattei", Milan.
- Spagnolo, G. (2000b), "How the law solves Bertrand's Paradox and enforces collusion in auctions" revised version of *F.E.E.M. Nota di Lavoro No. 52.00*, Fondazione ENI "Enrico Mattei", Milan.
- Spagnolo, G. (2004), "Divide et impera: Optimal leniency programs", *CEPR Discussion Paper*, No. 4840.
- Spagnolo, G (2005), "Cartels criminalization and their internal organization", in K.J. Cseres, M.P. Schinkel and F.O.W. Vogelaar (eds), *Remedies and Sanctions in Competition Policy: Economic and Legal Implications of the Tendency to Criminalize Antitrust Enforcement in the EU Member States*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing
- Spagnolo, G. (2008), "Leniency and whistleblowers in antitrust", in P. Buccirossi (ed.), *Handbook of Antitrust Economics*, Cambridge, MA: MIT Press.
- Stigler, G. (1964), "A theory of oligopoly", *The Journal of Political Economy*, **72**, 44–61.
- Uytsel, S. (2015), "Anti-cartel enforcement in Japan: Does leniency make the difference?" in C. Beaton-Wells and C. Tan (eds), *Anti-Cartel Enforcement in a Contemporary Age: Leniency Religion*, Oxford: Hart Publishing.
- Veljanovski, C. (2010), "European Commission cartel prosecutions and fines, 1998–2006: An updated statistical analysis of fines under the 1998 Penalty Guidelines", *Working Paper*, Case Associates/Institute of Economic Affairs.
- Wandschneider, F. (2014), "An experimental study of ringleader exclusion from leniency programmes", PhD thesis, University of East Anglia.
- Werden, G., S. Hammond and B. Barnett (2011), "Recidivism eliminated: Cartel enforcement in the United States since 1999", paper presented at Georgetown Global Antitrust Enforcement Symposium, Washington, DC, 22 September.
- Wils, W. (2016), "The use of leniency in EU cartel enforcement: An assessment after twenty years", *World Competition*, **39**(3), 327–88.
- Yeremin, D., A. Subbot and M. Mouradov (2011), "Russia", in S.J. Mobley and R. Denton (eds), *Global Cartels Handbook: Leniency: Policy and Procedure*, Oxford: Oxford University Press.
- Yusupova, G. (2013), "Leniency program and cartel deterrence in Russia: Effects assessment", *Higher School of Economics Research Program*, No. WP BRP 06/PA/2012.
- Zhou, J. (2013) "Evaluating leniency with missing information on undetected cartels: Exploring time-varying policy impacts on cartel duration", *TILEC Discussion Paper*, No. 2011-042.
- Zhou, J. (2016), "The dynamics of leniency application and cartel enforcement spillovers", *TILEC Discussion Paper*, No. 2016-006.



5. Assessing coordinated effects in merger cases

Natalia Fabra and Massimo Motta

1 INTRODUCTION

Merger control is one of the pillars of antitrust policy. It is necessary in order to ensure that anticompetitive mergers – that is, mergers that lead to a price increase, lower production, less variety, fewer innovations, etc. – do not take place. There are two mechanisms whereby mergers can give rise to anticompetitive effects: unilateral effects and coordinated effects.

The concept of *unilateral effects* refers to a situation where the merger allows the merging firms to unilaterally – that is, independently of the reaction of the remaining competing firms – increase their market power: because of the lower competitive constraints (a merger reduces the number of independent competitors in the industry), firms that would not have increased their prices (or reduced production, etc.) after their merger may find it profitable to increase them, even if all other firms' prices remained unchanged.¹

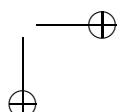
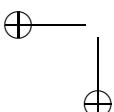
The concept of *coordinated effects* refers instead to the fact that after the merger it will become more likely that the merging firms *and* (at least an important subset of) their rivals will increase their market power by coordinating their actions. In other words, the term “coordinated effects” indicates the higher probability that after the merger the main firms in the market will reach a (tacit or explicit) collusive outcome or – if collusion was already taking place – would strengthen such an outcome, for instance by managing to reach higher collusive prices, or by making collusion more stable.

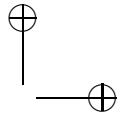
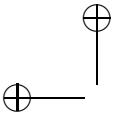
This chapter deals with coordinated effects of mergers, and its objective is to explain what they are and how they can be identified. Since a full understanding of collusion is fundamental to explaining coordinated effects, in the first part of this chapter we draw on the theoretical and empirical economic literature to answer two basic questions: what is collusion and what facilitates it? The analysis of collusion and of the factors that facilitate it is the building block for the analysis of coordinated effects in mergers, and provides us with important hints on how to conduct such analysis in *practice*.

Whenever an agency is facing a merger, it will have to make an analysis of the market, to gather clues as to whether the merger may raise unilateral effects, or coordinated effects, or whether it raises no danger of increased market power.² When conducting such an analysis, some clues to whether coordinated effects may be relevant at all could be obtained by looking at very *simple indicators*. In our opinion, two will be especially important. The first concerns market structure: tacit collusion is unlikely to arise unless post-merger there are only two or

¹ It is important to notice though, that *after* the merging firms increase their prices (or reduce their output), the rival firms will modify their decisions in turn. But the overall effect will generally be anticompetitive.

² In equilibrium, a merger might give rise to either unilateral effects or coordinated effects, not both at the same time. However, an antitrust authority should assess both, because it may not be clear *a priori* which one would arise at equilibrium (finding that a collusive equilibrium is more likely to occur after the merger does not mean it will occur with probability 1). Note also that the unilateral effects analysis will generally give the lower bound to the potential price increase of the merger.





92 *Handbook of game theory and industrial organization: applications*

three firms in the market, with considerable symmetries among them. The second concerns past history of collusion: a motivated suspicion of a strengthening of coordinated effects should arise if the industry has a past history of collusion, with firms having developed a web of relationships (joint ventures, purchasing and/or distribution agreements, cross-directorates etc.), or a system of exchange of information (or other price schemes that improve monitoring), or if suspiciously parallel price movements have taken place over time. In the second part of this chapter we review these as well as other “screens” and indexes one may want to look at in order to assess the likelihood of coordinated effects in practice.

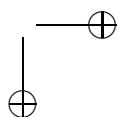
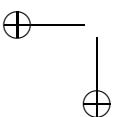
Since we believe that the analysis of past merger policy is fundamental to better enforcing competition policy, we devote the last part of the chapter to the issue of how coordinated effects have been applied in European merger control. Originally, EU Merger Regulation 4064/89 stated that mergers that would *create or strengthen a dominant position* (defined as the ability to behave to an appreciable extent independently of rivals and customers – and effectively amounting to the possession of very large market power) would be declared incompatible with the common market. However, the European Commission soon realized that there were mergers that appeared to be anticompetitive even if they did not give rise to a (single-firm) dominant position. To cope with such situations, the European Commission borrowed from the existing jurisprudence the concept of *collective dominant position* (or joint dominance), which was then applied to several cases (in an increasingly extensive manner), and which was arguably used as a way to address possible anticompetitive situations, perhaps also beyond the concept of coordinated effects. The review of the recent *AGF/GBI* merger case helps illustrate how the Commission applies the Guidelines to assess coordinated effects in merger cases.

The chapter continues in the following way. Section 2 describes the potential anticompetitive effects of horizontal mergers: unilateral and coordinated effects. It also addresses the main questions that need to be explored in an assessment of coordinated effects: whether collusion in the ex post merger market would be sustainable (enforcement problem), whether firms would be able to reach a mutual understanding or agreement (coordination problem), and whether the merger would relax both problems, thus facilitating collusion. Section 3 reviews some approaches that should help identify coordinated effects and “quantify” their relevance in practice. Section 4 describes the evolution of the policy on coordinated effects in European merger control, and discusses the *AGF/GBI* case. Section 5 of the chapter concludes.

2 UNDERSTANDING COLLUSION TO BRING A COORDINATED EFFECTS CASE

A merger between competitors (known as a horizontal merger)³ might give rise to an increase in prices and thus be anticompetitive. This might be due to two distinct effects: unilateral and coordinated effects.

³ Unless explicitly mentioned, throughout the report we focus on horizontal mergers among producers. Similar principles also apply to horizontal mergers among buyers, who have an incentive to reduce demand and lower prices. However, mergers among buyers can lead to a distinctive feature, namely, buying power, whose impact on coordinated effects is discussed in Section 2.2. Coordinated effects in vertical merger cases are discussed in the Section 2.3.1.



To illustrate these effects, consider a set of single-product firms selling substitute products. An increase in the price of one product translates into an increase in the sales of another. However, this positive externality is not taken into account by firms when setting their prices given that the increase in sales benefits rival firms. A merger between two firms would allow them to internalize such externality and, absent any cost synergies,⁴ would induce them to push prices up. This holds true regardless of the reaction of the outsiders. If such firms optimally react by also increasing their prices, the *unilateral effects* of the merger would be enhanced.⁵ This leads to a new outcome in which all firms end up charging higher prices than before the merger, with the merger firm charging relatively higher prices than the non-merged firms.

Firms could also sustain higher prices after a merger by coordinating their actions. A merger leads to *coordinated effects* if it makes it more likely that the merging firms *and* (at least an important subset of) their rivals increase their market power through coordination. In other words, the term “coordinated effects” indicates the higher probability that after the merger the main firms in the market will reach or strengthen a (tacit or explicit) collusive outcome.

To assess whether a merger would create coordinated effects, one should address the following three questions:⁶

1. Would collusion post-merger be possible and sustainable? [*enforcement problem*]
2. Would firms be able to reach a collusive agreement and adapt it to the possibly changing market conditions? [*coordination problem*]
3. Would the merger enhance the likelihood of collusion? [*coordinated effects*]

The first question refers to the *enforcement problem*: for collusion to be sustainable, firms must find it in their own interest to respect the collusive agreement. The stability of collusion in the ex post merger market is therefore a necessary condition for the merger to give rise to coordinated effects.

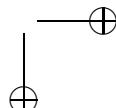
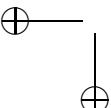
However, it is not sufficient: the fact that firms could sustain collusion does not mean that they actually succeed in doing so.⁷ For the market outcome to be collusive, it is also necessary that firms solve a *coordination problem*, i.e., they have to agree on which strategy to follow, which price they want to set or which level of output they want to produce, how they will adapt it to changes in the market environment, among many other dimensions of the agreement. The coordination problem might be particularly acute when firms are asymmetric or when they

⁴ Horizontal mergers can also generate efficiency gains. If such gains are sufficiently strong, they might offset the anticompetitive effects of mergers. See Farrell and Shapiro (1990) for a formal analysis. In Section 2.2 we also discuss the effect of cost asymmetries on coordinated effects.

⁵ This effect is shared by all models with “strategic complements,” e.g., in which the marginal profit of increasing one’s price is higher the higher the price charged by the other firms. This does not hold true in the presence of “strategic substitutes,” e.g., when the marginal profit of increasing one’s quantity is higher the lower the quantity produced by the other firms. In particular, when firms compete by choosing output, the outsiders react by expanding their output after the merger. However, the overall effect of the merger is an output contraction, given that the merging firms’ output reduction is stronger than the outsiders’ output expansion.

⁶ In line with this approach, the EU Horizontal Merger Guidelines (HMGs) (2004) state that “[t]he Commission examines whether it would be possible to reach terms of coordination and whether the coordination is likely to be sustainable. In this respect, the Commission considers the changes that the merger brings about” (para. 42).

⁷ Even when collusion is sustainable, there are typically many outcomes that firms could end up reaching, which involve lower equilibrium profits, e.g., the equilibrium at the competitive benchmark. Firms might also have conflicting interests as to which equilibrium to play, or as to how to adapt it to changing market conditions.



sell differentiated products, as such features may give rise to a conflict of interests among them. Instead, communication among firms might allow firms to more effectively solve the coordination problem. These issues are addressed by the second question above.⁸

There is often a positive link between the circumstances that make collusion more easily enforceable, and those that facilitate coordination on a collusive equilibrium.⁹ For instance, as we discuss below, enforcing collusion and coordinating in a collusive equilibrium is easier the smaller the number of firms. However, enforceability does not imply coordination, or vice versa, i.e., there might be contexts in which coordination is possible and yet collusion is not enforceable, or vice versa.

Answering the first two questions allows one to assess whether the merger would give rise to coordinated effects. On the one hand, one could argue that the sustainability of collusion and firms' ability to coordinate on a collusive equilibrium are not sufficient to prohibit a merger on the basis of coordinated effects. For instance, if firms already collude in the pre-merger market structure, one could be tempted to conclude that the merger does not have any incremental effect on collusion. However, whereas this might be a possibility in economic models under particular assumptions,¹⁰ it is unlikely to hold in practice. If collusion took place before the merger, most likely the merger will enhance it, by making it more stable (there would be a lower risk that a shock might result in a breakdown of collusion) or permitting firms to reach higher prices among the sustainable collusive ones. Therefore, if the first two questions indicate evidence of collusion before the merger takes place, then the merger should not be allowed on the basis of coordinated effects.¹¹

In the next sections we first define the term “collusion” and describe the mechanisms by which firms can make it sustainable over time. We then examine the factors that facilitate collusion by relaxing the enforcement and the coordination problems. Last, we turn to the issues that need to be examined when evaluating the coordinated effects of horizontal mergers.

2.1 What is Collusion?

2.1.1 Tacit versus explicit collusion

For economists, collusion arises when firms are able to sustain prices above some competitive benchmark,¹² in the fear that deviations from the agreed behavior would trigger periods of intense rivalry. Thus, economists put the emphasis on the market outcome and the incentive structure supporting it, regardless of whether firms achieve such an outcome through either

⁸ While policy discussions tend to put most emphasis on the coordination problem, the standard modeling approach focuses on the enforcement problem. Indeed, economic theory provides many insights on the nature of collusive equilibria, but says little on how firms coordinate (or not) on a particular collusive equilibrium, and on which one. There are some recent exceptions. See Harrington (2012b) and Lu and Wright (2010) for analyses on how firms reach a mutual understanding through price leadership and price matching. See also discussion in Section 3.3.

⁹ As Harrington (2012a) notes: “conditions for a firm to optimally initiate collusion are, to some degree, dual to the conditions for a firm to optimally sustain collusion.”

¹⁰ For instance, if the discount factor is very close to one and if coordination problems are assumed away, then collusion on the monopoly outcome will be possible regardless of the number of firms.

¹¹ Perhaps the only caveat in this respect is a *de minimis* argument. Indeed, one might argue that a merger between two small competitors is unlikely to further enhance coordination even in markets in which collusion was already sustainable. Even in this case, though, one may object that allowing a merger between small firms may lead to other such mergers that would eventually result in a much more concentrated industry.

¹² It is important to stress that the competitive price may already incorporate a mark-up over marginal costs. This is the case in all oligopolistic models, except for the Bertrand model of price competition and perfectly homogeneous goods. Hence, prices above marginal costs do not necessarily reflect a collusive outcome.

tacit or explicit collusion. Instead, lawyers, judges, and antitrust authorities are concerned about the means by which firms reach and sustain a collusive outcome. As Joseph Harrington puts it, “there is a gap between antitrust practice – which distinguishes explicit and tacit collusion – and economic theory – which (generally) does not.”¹³

In most jurisdictions, only explicit agreements, for which there is hard evidence of communication, are considered illegal. In contrast, tacit collusion is generally not considered as a violation of antitrust law.¹⁴ However, both explicit and tacit collusion are taken into account when assessing the coordinated effects of horizontal mergers. Indeed, a merger might potentially facilitate cartel formation as well as give rise to conditions that relax the enforcement problem faced by firms when colluding either explicitly or tacitly. Accordingly, the assessment of coordinated effects through merger control can constitute a powerful ex ante tool to deter cartel formation as well to fight tacit collusion. The latter is particularly relevant given the difficulties in fighting *tacit* collusion ex post.

2.1.2 How can firms sustain collusion?

Both theory and experience suggest that frequent interaction among firms may have a dramatic effect on market performance: in a dynamic setting, firms may learn to coordinate their strategies, and hence compete less aggressively with each other over time, through either tacit or explicit agreements. However, colluding is not an easy task as each firm is tempted to cheat on the tacit agreement. This is true even when firms collude explicitly, given that if one firm does not comply with the agreement, such a firm can clearly not be taken to court for breach of contract by the other cartel members.¹⁵

To illustrate the incentives faced by colluding firms, let us consider a simple set-up. Suppose that all firms in the market sell their products at a price above the competitive price as they understand that it is in their common interest to do so. Knowing that all other firms are setting a high price, any firm could profitably deviate by undercutting it, as the firm would increase its sales with only a slight price reduction. So, what discourages firms from undercutting each other? It is the fear that the rivals will react by setting very low prices as soon as they detect a price cut. In other words, the fear that the price deviation will trigger periods of intense rivalry is the disciplining device that makes firms overcome their short-run temptation to deviate, and allows them to sustain collusive outcomes.¹⁶

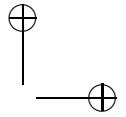
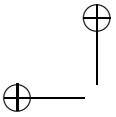
In order to sustain collusion, it is necessary that firms are able to detect deviations, for which they need to monitor each other. Equally critical for the sustainability of collusion, is firms’ ability to credibly retaliate when they detect a deviation. But the possibility of inflicting strong punishments has to be assessed relative to the gains from deviation. Indeed, colluding firms

¹³ See Harrington (2005). For a discussion on the distinction between the economic and legal approaches to collusion, see, for instance, Kaplow and Shapiro (2007).

¹⁴ See Motta (2004) and Mezzanotte (2009) for a discussion.

¹⁵ In the presence of leniency programs, the deviant would be the one to denounce the cartel to the antitrust authority. After the deviation the cartel would in any case destabilize, but thanks to the leniency application the deviant would benefit from a reduced fine or even amnesty. For this reason, leniency programs hinder collusion. See Motta and Polo (2002).

¹⁶ This explains why collusion can only be reached in dynamic settings (i.e., when firms interact repeatedly): in static settings the reduction in future profits cannot be used as a credible threat to discourage deviations simply because the future does not exist. Nevertheless, repeated interaction is not sufficient: interaction has to be infinite, or at least for an undetermined number of periods. Otherwise, in the last period all firms would deviate knowing that future punishments are not feasible. In turn, this makes it impossible to threaten firms in previous periods, so that collusion unravels in all periods.



face a trade-off. On the one hand, if a firm respects the collusive agreement, it gets collusive profits in the current period as well as in all future periods. On the other hand, if it deviates, it gets a higher profit in the current period, but much lower profits in the future as the deviant will be punished. Collusion will thus be sustainable if the value of current and future collusive profits exceed the value of current deviation profits followed by the flow of future punishment profits. This trade-off involves current short-run gains versus future losses. Therefore, any factor that enhances the future losses from deviating or that mitigates the current short-run gains from deviation will tend to facilitate collusion. We expand on this in the next section.

2.2 Which Factors Facilitate Collusion?

A factor facilitates collusion if it allows firms to sustain and to agree on a collusive strategy in markets where collusion would otherwise not be sustainable. A facilitating practice may also strengthen collusion, by allowing firms to raise the profitability of the collusive agreement in markets in which firms were already sustaining prices above the competitive benchmark.

A correct identification of the factors that facilitate collusion is particularly relevant in merger analysis as it is in those industries more vulnerable to collusion where the coordinated effects of mergers are more likely to arise. The section on coordinated effects of the EU Horizontal Merger Guidelines (HMGs) (2004) starts by noting that “[i]n some markets the structure may be such that firms would consider it possible, economically rational, and hence preferable, to adopt on a sustainable basis a course of action on the market aimed at selling at increased prices” (para. 39). The aim of this section is to identify the factors that make some markets particularly more prone to collusion than others.

A factor facilitates collusion if (i) it relaxes the conditions that guarantee that firms have no incentives to deviate from the collusive agreement (*enforcement problem*);¹⁷ or if (ii) it facilitates coordination on a collusive equilibrium (*coordination problem*). The first condition is met if collusive profits increase, deviation profits are reduced, or if the punishment threat becomes more severe. An improvement in monitoring, so that deviations can be more quickly and more accurately detected, would also relax the enforcement problem and thus facilitate collusion. The second condition is met when firms’ conflict of interests is mitigated, or when they can more effectively communicate to coordinate their actions.

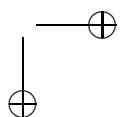
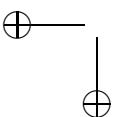
For ease of exposition, we classify the factors that affect collusion under four broad categories: (a) supply factors; (b) demand factors; (c) transparency, communication and information exchange; and (d) corporate governance structures.

2.2.1 Supply factors

Number of firms The number of firms in the market plays a crucial role in determining the likelihood of collusion. As expressed in the 2004 EU HMGs, “it is easier to coordinate among a few players than among many.” In other words, a small number of competitors find it easier to overcome the coordination problem.¹⁸ Furthermore, once firms have reached a consensus on the collusive agreement, it is the easier for them to sustain collusion the fewer they are. That

¹⁷ In economic theory, these are referred to as incentive compatibility constraints.

¹⁸ The idea that coordination is easier the smaller the number of firms is intuitive, but there is little economic literature on this result. See Compte and Jehiel (2010). Huck, Normann and Oechssler (2004) and Engel (2007) provide experimental evidence in the lab supporting this result.



is, a small number of competitors also find it easier to overcome the enforcement problem: first, the smaller the number of firms in the industry the easier it is to monitor each other; and second, the temptation to deviate from the collusive agreement is also weaker since collusive profits have to be shared among fewer firms.

Entry The number of firms in an industry can increase through entry. As acknowledged by the 2004 EU HMGs, one of the conditions for the sustainability of collusion is that “the reactions of outsiders, such as current and future competitors not participating in the coordination . . . should not be able to jeopardize the results expected from the coordination” (para. 41). Indeed, in industries with low barriers of entry, firms will find it difficult to sustain collusive agreements.¹⁹ This holds true regardless of how the entrant behaves and how the incumbents react to entry.²⁰

Excess capacity The degree of firms’ excess capacity is a key ingredient affecting collusion possibilities. When firms are capacity constrained, capacity constraints affect the size of the market that a firm can capture for itself when it deviates. Hence, the larger the firm’s unused capacity, the greater its incentives to deviate. However, capacity constraints also affect the scope of other firms to flood the market in order to reduce profits following a deviation. Hence, the larger the degree of excess capacity in the industry, the more effective is such a disciplining device. Since these two forces move in opposite directions, it is *a priori* not possible to conclude whether larger capacities at the industry level facilitate or hinder collusion.

Size asymmetries Let us start by considering a market made of symmetric firms, in the sense that they all sell homogeneous products that they can produce at equal costs. Any move away from symmetric market shares (which would raise concentration) would also hinder collusion. This is so since the firm with the small market share has more to gain by deviating and less to lose from being punished.

Differences among firms – such as differences in their productive capacities, in the features of their products, in the size and content of their product portfolios, or in their production costs – typically explain why market shares are asymmetric. The question is then: how do such fundamental asymmetries, which often translate into asymmetric market shares, affect collusion?

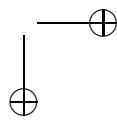
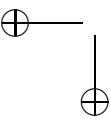
While asymmetries might have a different impact on the mechanisms affecting the incentives to collude, there is a robust result that says that firms’ asymmetries hinder collusion. Indeed, as we describe below, firm symmetry facilitates both the enforcement as well as the coordination problem.

Firm symmetry relaxes the enforcement problem, for one key reason: the scope of collusion is determined by the firm facing the greatest difficulties to collude (be it the large, or the small firm);²¹ as firms become more symmetric, there is a transfer in the ability to collude from those that find it easier to collude to those that face the greatest difficulties in colluding. This rebalancing in the incentives to collude unambiguously facilitates collusion.

¹⁹ Very often entry occurs in industries in which future demand is growing, and these two facts might affect collusion in opposite directions. See below for a discussion of collusion under demand fluctuations.

²⁰ The lysine price-fixing cartel in the mid-1990s provides an example of how incumbent firms react to entry.

²¹ Technically speaking, this is the firm whose incentive compatibility constraint is binding.



To fix ideas, consider a context in which market share asymmetries derive from differences in firms' product lines (Kühn, 2004 and Motta, 2004). If the size of a firm is a function of the number of product varieties it holds, then it is the small firm that faces the greatest difficulties in colluding.²² For a large firm, a reduction in the price of one of its varieties has a negative effect on the profits it makes through its other varieties. Hence, a large firm has a weaker incentive to deviate compared to a single-product firm, since the latter does not internalize the negative impact of a price cut on other varieties. Similarly, low prices after a deviation hurt the large firms relatively more than the small firm, and so the large firms' ability to hurt the small one is limited.

When market share asymmetries derive from capacity asymmetries, the mechanisms sustaining collusion differ from the one just described. Let us consider a model in which firms sell homogeneous products but are subject to asymmetric capacity constraints (Compte, Jenny and Rey, 2002). The large firm, and not the small one, is now the one that would benefit most from deviating, given that it could capture a greater fraction of the market were it to undercut the collusive price. Furthermore, the small firms cannot inflict strong punishments on the large firm given that, even when operating at full capacity, the residual demand left for the large firm would still be significant. Hence, the bigger the large firm the more difficult it is to discourage such a firm from deviating. A more equal distribution of firms' capacities would realign their incentives to collude and their capacity to punish deviators, thus facilitating collusion. In general, this implies that capacity asymmetries hinder collusion. Still, differences in concentration due to differences in the size of the small competitors should have no impact on the sustainability of collusion, as long as the size of the large firm remains unchanged.

When assessing the role of firms' asymmetries, it is also equally important to understand how they affect the coordination problem. According to the 2004 EU HMGs, "[c]oordinating firms should have similar views regarding which actions would be considered to be in accordance with the aligned behavior and which actions would not" (para. 44) and "[f]irms may find it easier to reach a common understanding on the terms of coordination if they are relatively symmetric, especially in terms of cost structures, market shares, capacity levels and levels of vertical integration" (para. 48). In other words, symmetry is generally assumed to relax the coordination problem.

When firms are engaged in tacit collusion, identifying a "focal point" in terms of prices or market shares may become less obvious the more asymmetric firms are. When firms sell homogeneous products and face equal costs of production, there is a single monopoly price that all firms should be able to compute, as they all share equal information. However, when their costs or the features of their products differ, agreeing on a common collusive price might not be an easy task, and firms might face conflicting interests as to which price to select. For instance, under cost asymmetries, low cost firms may prefer to collude on lower prices than high cost firms, and successful collusion might be preceded by periods of trial and error through prices until firms achieve a tacit agreement on a given price.²³

²² A similar result also arises in models with asymmetric capacities, which give rise to cost asymmetries (Vasconcelos, 2005).

²³ Phillips, Mason and Nowell (1992) provide experimental evidence showing that cooperation is more likely among firms with symmetric costs.

When firms are engaged in explicit collusion, bargaining can lead to efficient outcomes even among asymmetric firms.²⁴ However, inefficiencies might arise whenever firms' asymmetries are private information (e.g., firms do not know each others' costs, the features of their rivals' products, etc.). Therefore, to the extent that firms' asymmetries go hand in hand with asymmetric information, it is reasonable to expect that such asymmetries might hinder coordination on a efficient outcome.

Cost asymmetries Cost asymmetries also hinder collusion. In this case, the low cost firm, which is typically also the large firm, finds it more tempting to deviate from the collusive agreement: it has more to gain by deviating as at any price its markup is higher, and it fears the punishment that can be inflicted by its high-cost rivals less.²⁵

Matters are more complex when firms do not know each other's costs. Athey and Bagwell (2001)²⁶ analyze a model of collusion with private cost information in which firms might face independent cost realizations in every period. They show that successful collusion among firms with asymmetric costs might sometimes entail productive inefficiencies: a high cost firm must be given incentives to report its true cost, and such incentives may require that the high cost firm serves an inefficiently large share of the market.²⁷ Such productive inefficiencies hinder collusion as they reduce collusive profits. Incentives for the high cost firm to truthfully report its cost might also come through side-payments by the low cost firm, but these would leave traces of explicit collusion and cartel firms would thus risk being detected and fined.

Multi-market contact The possibility to sustain collusion might also depend on the number of markets in which the same set of firms interact; this is referred to as multi-market contact.^{28,29} Building on the intuition described above on the effects of asymmetries, pooling the incentives to sustain collusion across asymmetric markets can help mitigate asymmetries within markets. Furthermore, multi-market contact facilitates collusion through increases in the frequency of interaction.

2.2.2 Demand factors

Demand movements The sustainability of collusion is affected by demand movements over time. Consider first the case of a market whose demand is known to steadily grow over time. Collusion in this market is more easily sustainable than if the demand is decreasing for a

²⁴ Indeed, in bargaining models with sequential offers (Rubinstein, 1982), agreement is efficient as otherwise firms would continue bargaining until all efficiency improvements become exhausted.

²⁵ In contrast to this result, Miklos-Thal (2009) finds that, if side-payments are allowed, cost asymmetries facilitate collusion.

²⁶ See also Athey, Bagwell and Sanchirico (2004), who assume persistence in cost shocks instead of assuming that cost shocks are independent across periods.

²⁷ Athey and Bagwell (2001) also show that if firms are sufficiently patient, perfect collusion can be achieved without sacrificing productive efficiency. This can be achieved by promising a high cost firm today with a higher market share in a future period in which both firms have equal costs. Market transfers so achieved are sufficient to ensure truth-telling as long as the discount factor is sufficiently high.

²⁸ For example, Bernheim and Whinston (1990) show theoretically that, in some cases, multi-market contact can improve firms' abilities to sustain high prices by pooling the incentive constraints that limit tacit collusion.

²⁹ See Phillips and Mason (1992) and Evans and Kessides (1994) for evidence of multi-market contact and collusion.

simple reason: future demand affects the losses from deviating, which are the greater the higher future demand.³⁰

The same logic extends to contexts in which demand moves cyclically over time, across booms when demand is rising and across recessions when it is declining. If one compares the sustainability of collusion across two periods of the cycle with equal demand, one in a boom and the other in a recession, the incentives to deviate are the same but the losses from deviating are greater in the former. Hence, the scope for collusion is greater during booms than during recessions (see Haltiwanger and Harrington, 1991).³¹

In contrast to our previous discussion, both the European Commission and the Court of First Instance (CFI) view demand growth as a factor-hindering collusion.³² We can think of two plausible explanations for this divergence: first, competition authorities and courts emphasize the role of demand growth on promoting entry (Vasconcelos, 2008); and second, they view demand growth as a source of demand instability, which – as discussed below – might jeopardize collusion sustainability.

Unexpected demand shocks When expected future demand is the same across all periods, so that the expected losses from deviating are also constant, unexpected positive shocks in demand can disrupt collusion by enhancing firms' current incentives to deviate (Rotemberg and Saloner, 1986). For this reason, even when demand shocks can be observed ex post, demand volatility hinders collusion.

Buying power Demand volatility can be exogenous, e.g., as in electricity markets, or endogenous, e.g., when it is driven by the demand of a big buyer that can decide how to schedule orders. Following the same logic as above, a big buyer is able to disrupt collusion by concentrating its purchases rather than scheduling frequent and regular orders (Snyder, 1996).³³ In this sense, buying power, which gives the buyer the ability to reduce the frequency of the interaction, hinders collusion. In line with this reasoning, the 2004 EU HMGs state that “if a market is characterized by infrequent, large volume orders, it may be difficult to establish a sufficiently severe deterrent mechanism” (para. 53). The 2010 US HMGs contain a similar statement: “A firm is more likely to be deterred from making competitive initiatives by whatever responses occur if sales are small and frequent rather than via occasional large and long-term contracts” (Section 7.2).

³⁰ This logic might nevertheless be reversed if future punishment profits also depend on the value of future demand and the impact of future demand movements is greater on punishment profits than on collusive profits. Fabra (2005) shows that collusion is more easily sustainable when demand declines if firms are subject to severe capacity constraints.

³¹ Nevertheless, the above discussion assumes that the market structure remains unchanged despite demand movements. However, in markets where entry barriers are not too high, this need not be an adequate assumption. Indeed, entry is more likely during booms, just as exit is more likely during busts. The question is thus whether the impact of such changes in market structure prevail over the impact of demand movements on collusion. See Lepore and Knittle (2010) for an extension of Fabra (2005) with endogenous capacity choices.

³² The decision of the *Airtours/First Choice* merger case illustrates this view, as the CFI argued that evidence of “strong growth” in demand would undermine attempts to collude. See Section 4.1.3 for a discussion.

³³ The practice of concentrating large-volume orders at infrequent times was, for instance, followed by the US government when it bought vaccines in bulk in order to undo collusion (Scherer, 1980). By buying in bulk, the government both increases the stakes of each procurement auction and reduces the frequency of such auctions, thus increasing the bidders' incentives to deviate and constraining their ability to punish each other in the near future.

Demand uncertainty Demand volatility very often goes hand in hand with demand uncertainty.³⁴ If demand changes over time and if such movements cannot be publicly observed, then firms might find it more difficult to monitor each other as a reduction in demand – which depresses all firms' sales – and can be wrongly confounded with a rival's price cut. In contrast, when market demand is stable, inferring deviations from publicly available data is easier than when the demand is volatile. We postpone the discussion of collusion when there is imperfect monitoring until Section 2.2.3, where we discuss the role of market transparency in facilitating collusion.

2.2.3 Transparency, communication and information exchange

In this section, we first discuss the importance of market transparency, which by increasing the observability of prices and quantities, improves monitoring. We then turn to the importance of communication in facilitating coordination among firms on a particular outcome. We emphasize the role and effects of different types of communication (whether it refers to future conduct or current and past data, whether it is public or private, and whether it includes detailed or aggregate data) on both the risk of collusion and the potential efficiency losses of banning communication.

Transparency In order to sustain collusion, it is necessary that firms are able to detect deviations, for which they need to monitor each other. Monitoring is thus a key ingredient of any collusive agreement. One can distinguish two features that characterize the effectiveness of monitoring: how long it takes firms to detect any potential deviation, and how precise is the information that firms receive on whether a deviation has indeed taken place. Monitoring is clearly the more effective the quicker it allows the detection of deviations and the more accurate it is in reporting whether a deviation has taken place. Transparency improves monitoring in these two dimensions.³⁵

In order to understand the role of transparency, let us consider the case in which market demand is uncertain and transaction prices cannot be publicly observed. Firms only see their own sales, but do not observe demand shocks. Firms cannot infer deviations from the data they observe, given that low sales can be due either to a low demand realization or to undercutting by the rival firm. If periods of low sales were not followed by a number of periods of intense rivalry or price wars, then firms would deviate knowing that they would go unpunished. Hence, in opaque markets, price wars are a disciplining device needed to avoid deviations, even when such deviations do not take place. Given that during price war periods firms make low profits, the profitability of collusion is lower in opaque than in transparent markets, as in the latter, price wars are not used in equilibrium.

Practices aimed at increasing transparency Given the importance of monitoring, competition policy should pay special attention to practices that help firms monitor each other's behavior. One example of such a practice is given by communication on past conduct, which is

³⁴ However, this is not necessarily always the case. For instance, demand can be perfectly observable and perfectly predictable, and yet it can change and be volatile over time.

³⁵ This is acknowledged in the HMGs both in Europe as well as in the USA. For instance, the US HMGs (2010) state that “[a] market typically is more vulnerable to coordinated conduct if each competitively important firm's significant competitive initiatives can be promptly and confidently observed by that firm's rivals. This is more likely to be the case if the terms offered to customers are relatively transparent” (Section 7.2).

discussed shortly. Other commercial and pricing practices also increase observability of firms' actions. For instance, collusion is more difficult when firms produce scores of *heterogeneous products*, both because they would have to keep track of prices of too many products (which makes sustainability more difficult) and because different products' prices are likely to be affected in a different way when shocks occur, which makes coordination more difficult. But if firms organize prices in very few and well-defined *price categories*, then both coordination and monitoring become much easier.

In the same vein, *resale price maintenance* (RPM)³⁶ helps collusion among suppliers. Indeed, as shown in Jullien and Rey (2007), RPM can facilitate collusion by making it easier for firms to monitor each other. To see why this is the case, consider a context in which downstream markets are subject to shocks on demand or retail costs that producers cannot observe. In the absence of RPM, downstream prices would reflect these shocks; for instance, if retailers' costs decrease, part of the cost reduction would optimally be passed through to retail prices. On the one hand, this allows firms to make higher collusive profits, thus discouraging deviations; on the other hand, it also makes it harder for firms to distinguish price cuts due to cost shocks, from price cuts due to deviations. RPM removes retail price flexibility, and thus has the opposite effects: lower collusive profits but more effective detection. The overall effect might seem ambiguous. However, in those cases in which RPM has no efficiency effects, we can be confident that if firms decide to adopt RPM it is because the pro-collusive effect dominates.

Communication In order to assess the role of communication and information exchange, it is first important to understand whether it makes any difference if firms communicate or not. In other words, does it make any difference whether firms collude tacitly or explicitly? On the one hand, through explicit collusion, firms might be able to reach and sustain outcomes they would not otherwise achieve. This is so since explicit communication facilitates agreement among the collusive firms, allows the tailoring of the pricing and sales policies to the specificities of each cartel member, makes it possible to adapt the collusive policies to changing market conditions, and allows firms to more effectively monitor each other's behavior. On the other hand, communication among cartel firms is costly, as it leaves trails that can then be used to detect the cartel.

Given the importance of communication, a powerful tool to fight collusion would be to prohibit communication among firms whenever such prohibition entails no efficiency losses, or rather, whenever the potential gains of deterring collusion exceed the potential efficiency losses of banning communication. For this reason, it is important to distinguish two types of communication. First, firms might communicate about their future intended conduct, e.g., planned production, prices, new product releases, capacity decisions, etc. This information is "soft" as it conveys intentions only, and cannot be verified by rival firms. Second, firms might communicate about current and past conduct, e.g., current and past sales, prices, product features, input prices, information about customers, etc. This information is "hard" as it can be verified, e.g., through invoices, customers' declarations, etc.

Communication about future conduct is important for sustaining collusion. On theory grounds, it is not straightforward to demonstrate that communication about future intentions

³⁶ Under RPM, retail prices are set by producers rather than by distributors.

helps sustaining collusion, as such communication has no commitment value.³⁷ Still, it can be a powerful tool for collusive purposes since it might facilitate coordination on a specific outcome, as explained below.

In many contexts, firms can sustain collusion on several prices but first need to coordinate on which price they will all choose. For instance, suppose that collusion at the monopoly price is sustainable and that products are perfect substitutes. Then, prices sufficiently close to the monopoly price should be equally sustainable too, as profits from deviating or colluding at such prices are roughly similar to when the monopoly price is chosen. However, not knowing whether rival firms plan to collude at the monopoly price or at prices arbitrarily close to it, firms face “strategic uncertainty”: if a firm sets the monopoly price but its rivals set a slightly lower price, the former will make zero profits and collusion could collapse. In light of this, firms may prefer to collude on prices below the highest sustainable price.³⁸ Communication about the price that firms plan to set mitigates strategic uncertainty, and thus facilitates collusion on higher prices.³⁹

However, not all announcements about future prices are harmful. When firms announce their sale prices to consumers, and they commit to serve consumers at those prices, transparency increases on the demand side and it favors “shopping around”: prospective customers are better informed on the possible deals, and they will tend – other things being equal – to buy from firms that offer lower prices. In turn, this will make the market more competitive.

It is true that when price announcements are public, prices would become transparent not only on the demand side but also on the sellers’ side. The latter effect would in principle favor collusion, but empirical evidence shows that it is the former effect that prevails.⁴⁰ It is important to stress, though, that for such a pro-competitive effect to take place, announcements should not only be public but also carry a commitment value towards consumers.⁴¹

Communication about firms’ future *production* plans is also unlikely to increase efficiency as it implies no commitment (plans can be changed), and it is unlikely to be informative to consumers. Instead, this type of information exchange may allow firms to reduce strategic uncertainty and thus to more effectively collude too. This example illustrates the practice

³⁷ In the jargon of economic theory, this is referred to as “cheap talk.”

³⁸ In games with multiple equilibria, one can apply the concept of *risk dominance* in order to select a plausible equilibrium (Harsanyi and Selten, 1988). In symmetric games (e.g., if symmetric firms charge the same price, they all get equal profits) this criterion allows for a simple interpretation: if firms are unsure about which price the rival will choose and assign equal probability to the rival choosing either a low or a high price, then the low price equilibrium risk dominates the high price equilibrium if the expected payoff from choosing the low price exceeds the expected payoff from choosing the high price. For instance, if firms consider choosing the monopoly price or one slightly below, choosing the latter is the risk-dominant equilibrium.

³⁹ The role of communication in eliminating strategic uncertainty has been explored in experimental settings. It has been shown that in the presence of strategic uncertainty, firms collude on prices below the monopoly level even when pricing at the monopoly level is also an equilibrium. See Cooper et al. (1989) and Van Huyck, Battalio and Beil (1990).

⁴⁰ See Motta (2004, pp. 152–156) for a discussion.

⁴¹ A famous case of communication of future intentions involved the Airline Tariff Publishing Company (ATP) in 1994, which used to collect and store data on airline fares quoted on computer reservation systems. Price announcements through ATP were public but had no commitment value towards consumers: airlines could enter future prices into the ATP system but could also change those prices before they could be effectively available for customers. Therefore, ATP constituted a pure vehicle for price coordination with no real price effects, very much as when firms are sitting around a table discussing future prices (US Department of Justice, 1994). In this case, whether potential buyers see the discussion or not, it makes little difference.

followed by the US automobile industry, which used to exchange production plans via the trade press (see Doyle and Snyder, 1999).

Communication about past conduct is also very important for sustaining collusion, though for different reasons. As argued above, the ability to monitor each other is crucial for the sustainability of collusion. Therefore, in markets in which firms cannot directly observe each other's price or output choices, communication about past conduct allows firms to overcome the lack of transparency. The more disaggregated the data (e.g., individual price choices and individual sales rather than average market price or aggregate sales) the more effective will communication be in allowing firms to detect deviations and to tailor punishments to the deviant.

2.2.4 Corporate and governance structure

Partial ownership arrangements (also referred to as cross-ownership) constitute passive investments as the acquiring firm gains no control over the decision taken by the firm whose stock it has acquired. Still, partial ownership arrangements may impact firms' conduct both in static as well as in dynamic games. In oligopolistic markets, when a firm increases its output it does not internalize the externality it imposes on others as the market price goes down. Hence, firms tends to over-produce above the level that maximizes industry profits. However, when holding shares of competitors, firms are able to at least partially internalize this negative externality, so that the market outcomes approach the monopoly outcome even in a static setting.⁴² In the limiting (though probably unrealistic case) in which firms retain control but exchange their stock across them, the monopoly outcome can be achieved with no need to collude.

Partial ownership arrangements also change firms' incentives to sustain collusive outcomes.⁴³ Authorities typically view cross-ownership as a factor facilitating collusion. For instance, the EU HMGs (2004) note that “[s]tructural links such as cross-shareholding or participation in joint ventures may also help in aligning incentives among the coordinating firms” (para. 48). Indeed, under cross-ownership deviation incentives are mitigated, given that a deviation by one firm imposes losses on others. Hence, cross-ownership facilitates collusion. Like *cross-ownership*, *cross-directorships* and *joint ventures* may also offer opportunities for competitors to talk to each other, thereby making coordination easier. Similarly, purchasing and/or distribution agreements can also serve the same purpose.

2.3 Is There a Coordinated Effect?

Putting together the above insights, when would the merger make collusion easier, more stable, more effective, and when would the mechanisms to sustain it be more easily agreed upon after the merger? If, in the light of the analysis developed in the previous section, collusion was already sustainable before the merger, it is highly likely that the merger would further strengthen firms' coordination. Hence, the analysis of whether the merger would create coordinated effects need not go much further. However, in those markets in which collusion

⁴² For instance, in January 2011, the UK Office of Fair Trading (OFT) opened an investigation into Ryanair's minority stake in Aer Lingus because it believed that it potentially raised competition concerns. The OFT press release can be found at <http://webarchive.nationalarchives.gov.uk/20140402184437/http://www.oft.gov.uk/news-and-updates/press/2011/01-11>.

⁴³ See Gilo, Moshe and Spiegel (2006) for an analysis of the effects of partial cross-ownership on the sustainability of tacit collusion. See also Buccirossi and Spagnolo (2007) for a discussion.

was not likely to be sustainable before the merger, one should conduct a careful analysis on the impacts of the merger on collusion.

The most straightforward effect of a merger is the reduction in the number of firms in the market. This alone has a direct effect on the incentives to collude: collusive profits have to be shared with fewer firms, so that the temptation to deviate from the collusive agreement is weaker. The reduction in the number of firms also creates unilateral effects, i.e., even in the absence of collusion, competition tends to be weaker the smaller the number of firms in the market.⁴⁴ While this might weaken the punishment threat, the deviation effect is of a higher order of magnitude than the punishment effect, implying that a reduction in the number of firms facilitates collusion despite the unilateral effects of the merger. The above, coupled with the fact that the reduction in the number of firms also relaxes the coordination problem (Section 2.2), unambiguously indicates that horizontal mergers facilitate collusion. However, this should not be misinterpreted to conclude that all mergers make collusion sustainable, as other factors also have to be assessed.

Among other relevant factors, it is particularly important to assess the effect of mergers on market structure; in particular, whether market structure becomes more or less symmetric after the merger.⁴⁵ As discussed in Section 2.2 above, mergers that make the large firm smaller or the small firm larger (i.e., symmetry increasing mergers) tend to facilitate collusion by relaxing the enforcement problem. Intuition also suggests that symmetry facilitates coordination on a collusive outcome. Hence, even if a merger involves a reduction in the number of firms, it might hinder collusion if it increases asymmetries among firms.

If there are any concerns that a merger would lead to coordinated effects, remedies should involve divestments that increase asymmetries among existing firms. A highly illustrative merger case in this respect is the *Nestlé/Perrier* case (see Section 4.1.2).

While horizontal mergers may weaken competition, they can also induce important *efficiency gains*. Indeed, if efficiency gains are sufficiently large, they may offset the otherwise negative effects of mergers on overall welfare. This question is well understood when it comes to assessing the trade off between efficiency gains and unilateral effects,⁴⁶ but much less attention has been devoted to the analysis of the interaction between efficiency gains and coordinated effects. Still, the discussion of cost asymmetries in Section 2.2 can shed some light on this issue: efficiency gains by the merging firms enhance cost asymmetries, which in turn hinders collusion. Furthermore, even if collusion is still sustainable after the merger, efficiency gains may imply an output transfer from the less efficient to the more efficient firms, as well as a reduction in the collusive price. Assessing the trade-off between efficiency gains and coordinated effects is nevertheless a difficult task: not only do prospective efficiency gains have to be estimated (as in a unilateral effects case), but also the impact of such gains on the likelihood of collusion.

Mergers can also affect the sustainability of collusion through its effects on multi-market contact among firms.⁴⁷ The idea is that collusion in all markets can be facilitated if mergers

⁴⁴ For instance, this is true in a Cournot model, when firms compete by choosing quantities.

⁴⁵ See Fonseca and Normann (2008) for experimental evidence of the effects of asymmetric mergers on collusion.

⁴⁶ See Whinston (2006), Motta (2004) and Motta and Tarantino (2016).

⁴⁷ Issues of multi-market contact have recently been raised in European merger cases. In 2007, Elopak and SIG, which were the main competitors of Tetrapak in the aseptic and fresh carton markets respectively, planned to merge. The Commission opened an in-depth investigation, but it was closed because the merger bid itself failed in the face of an alternative bidder. See Kühn (2008) for a discussion. See also Montero and Johnson (2012) for a recent theoretical analysis.

make the market position of firms across such markets more symmetric. To illustrate this, let us go back to the example used before: consider two markets, A and B; firm 1 is present in both markets, while firms 2 and 3 are only present in market A and B respectively. In market A firm 1's market share is s and firm 2's is $1 - s$; while in market B firm 1's market share is $1 - s$ and firm 3's is s . A merger between firms 2 and 3 creates multi-market contact between firm 1 and the new merged entity, and this implies that firms become symmetric across markets. Whereas before the merger within market share asymmetries would make collusion difficult, the merger facilitates collusion by making firms symmetric. While this example illustrates a concentration between two firms in unrelated markets, i.e., a conglomerate merger, the intuition extends to horizontal mergers with conglomerate aspects.

The structure of cross-ownerships among merging firms also has to be carefully assessed in a coordinated effects analysis. Consider again a simple example. Suppose that firm 1 owns a certain number of shares of firm 2, while firm 2 owns the same number of shares of firm 3. The latter is the one that finds it more difficult to collude, given that the other two firms' incentives to deviate are tempered by the fact that a deviation hurts them indirectly through their partial ownership of rival firms. A merger between firms 2 and 3 would imply that all firms in the market have fully symmetric cross-ownership of one another, thus facilitating collusion.

2.3.1 Coordinated effects of vertical mergers

Just as horizontal mergers have the potential to facilitate collusion, so do vertical mergers. This can be due to some of the effects highlighted before when assessing the coordinated effects of horizontal mergers. For instance, a vertical merger might make active firms more symmetric if after the merger all firms are vertically integrated and therefore share the same type of production (and distribution) costs. In turn, this would facilitate collusion.

In this section we focus on the coordinated effects that arise only because of the vertical relationship. As shown by Nocke and White (2007),⁴⁸ vertical mergers might facilitate collusion among producers. On the one hand, when two firms vertically integrate, the size of the downstream market that a deviant can capture is smaller, given that the integrated retailer is loyal to its upstream subsidiary. This effect, which is referred to as the *outlets effect*, reduces deviation profits and thus facilitates collusion. On the other hand, it is also more difficult to discipline a vertically integrated firm given that it benefits, in any event, from the profits made by its downstream subsidiary. This effect, which is referred to as the *punishment effect*, reduces the severity of the punishment threat and thus hinders collusion. However, the outlets effect dominates, implying that vertical mergers facilitate upstream firms' ability to collude.

We believe that this conclusion would be strengthened in markets with imperfect observability, e.g., because upstream producers cannot observe each other's prices and these cannot be inferred from retailers' price or output choices. Indeed, if the downstream market is subject to random shocks, producers cannot distinguish whether a price cut by a retailer is due to an adverse demand shock or to a deviation by an upstream rival (just as described in Jullien and Rey, 2007; see Section 2.2.3 above). In this context, vertical integration would allow the upstream producer to better monitor the behavior of its upstream rivals, given that its downstream subsidiary would have information on retail conditions. This concern is also contained in the 2008 EU Non-Horizontal Merger Guidelines (NMGs), which state that

⁴⁸ See also Normann (2009), which considers linear prices, and allows for raising rivals' costs effect.

“[v]ertical integration may give upstream producers control over final prices and thus monitor deviations more effectively” (para. 86). This effect, if combined with the *outlets effect* of vertical integration, would again point to the same conclusion: vertical mergers have the potential to facilitate upstream collusion.

This theory was to the best of our knowledge first adopted by the UK Competition Commission in the *Anglo American/Lafarge S.A.* case (2011). The merger (involving cement and concrete producers) did not create vertical integration, but increased it. According to the Competition Commission, it would have allowed Lafarge better access to information. Integration with Anglo American would in particular provide Lafarge with a better understanding (in terms of overall information and its geographic distribution) of the ready-to-mix (RMX) market. The ownership of the RMX plants would increase the knowledge of the local market conditions and allow better monitoring of deviations, whereas absent the merger, Lafarge would find it difficult – in areas where it does not have RMX plants – to understand whether lower sales would be due to an overall decline in demand or a deviation by competitor.

The above conclusion is also reflected in both the US and EU NHMGs; however, their reasoning is somewhat different. In particular, the NHMGs highlight the role of vertical integration in facilitating collusion through the elimination of “disruptive buyers.” For instance, the 1984 US NHMGs state that: “The elimination by vertical merger of a particularly disruptive buyer in a downstream market may facilitate collusion in the upstream market” (Section 4.222).⁴⁹ This concern rests on the following intuition: if sales to a disruptive buyer are relatively important, then upstream firms might have more incentives to deviate in order to secure business with such a relevant buyer. A merger with such a buyer reduces rivalry, and thus facilitates collusion.

Still, this result can also be accommodated within our previous reasoning. Note that if “sales to a particular buyer are sufficiently important,” such a buyer is necessarily a big one. Vertical integration with a big buyer enhances the outlets effects: the larger the integrated buyer, the smaller the fraction of the downstream market that the potential unintegrated upstream producers can capture if they deviate. Hence, a vertical merger with a big buyer facilitates collusion more than a vertical merger involving a relatively smaller retailer (Nocke and White, 2010).

3 QUANTIFYING COORDINATED EFFECTS CASE IN PRACTICE

3.1 Preliminary Considerations: HHI, Symmetry, and Past Collusion

The analysis of collusion and of the factors that facilitate it is the building block for the analysis of coordinated effects in mergers, and provides us with important hints on how to conduct such analysis in *practice*. Whenever an agency is facing a merger, it will have to make an analysis of the market, to gather hints as to whether the merger may raise unilateral effects, or coordinated effects, or whether it raises no danger of increased market power. When conducting such an analysis, some hints of whether coordinated effects may be relevant at all could be obtained by looking at very *simple indicators*.

⁴⁹ The 2008 EU NHMGs include similar concerns. See paragraph 90.

In our opinion, the following will be especially important. First, in general tacit collusion is unlikely to arise unless after the merger there will be two or three firms with a very important share of the market (say, more than 70 percent), and there will be considerable symmetry among them. This consideration is only partially aligned with what is probably considered the main indicator for anticompetitive mergers, that is, the Herfindahl-Hirschman Index (HHI) of industrial concentration.⁵⁰ Given that the HHI is the sum of the squared market shares, the index is the higher – other things being equal – the fewer the firms in the industry. However, the HHI decreases with symmetry. Therefore, we suggest that an agency should not only look at whether the industry is concentrated, but also – for the purpose of deciding whether to look into coordinated effects – if market shares (and capacities) are sufficiently symmetric across the main players.

Second, a motivated suspicion of strengthening of coordinated effects should arise whenever one discovers that the industry has a past history of collusion (for instance, cartels have been investigated following suspicious conduct, or successfully prosecuted, perhaps also in similar or adjacent markets), when firms have developed a web of relationships (joint ventures, purchasing and/or distribution agreements, cross-directorates etc.), when they have established a system of exchange of information (or other price schemes that improve monitoring), or when suspiciously parallel price movements have taken place over time (in this respect, we shall explain in Section 3.2 that there are a number of relative simple collusive “markers” or “screens” one may want to look at).

3.2 Screening for Coordinated Effects

The EU HMGs state that evidence of past coordination is particularly important when assessing the coordinated effects of mergers, particularly so if the characteristics of the relevant market have not changed significantly or are unlikely to change in the near future. Evidence of coordination in similar markets is equally relevant (para. 43). In line with this, the 2010 US HMGs state that “conditions are conducive to coordinated interaction if firms representing a substantial share in the relevant market appear to have previously engaged in express collusion affecting the relevant market.... Failed previous attempts at collusion in the relevant market suggest that successful collusion was difficult pre-merger but not so difficult as to deter attempts, and a merger may tend to make success more likely.” The view that firms that colluded in the past will try to do so again is supported by empirical evidence showing that cartel breakdown tends to be followed by attempts to re-establish cartels (Levenstein and Suslow, 2002).

Economic analysis can play a major role in screening, i.e., identifying those industries in which cartel formation and tacit collusion are more likely. Screening is the first step in the process of detecting cartels, and it may or may not end up in prosecution. Indeed, it is a useful tool in that it picks those industries where antitrust authorities should devote more effort to looking for collusive evidence (be it hard evidence, or competing explanations for observed behavior). Similar tools and indicators as the ones used for screening can also be useful for identifying those industries in which a merger would facilitate cartel formation or tacit collusion.

⁵⁰ See Coate (2005) for an empirical investigation of what are the main factors behind the US Federal Trade Commission’s decisions to challenge a merger. HHI levels and changes are definitely one of the variables with most explanatory power.

There are two main approaches for screening: the structural and the behavioral approach. The structural approach checks whether those factors that facilitate collusion, as reviewed in the previous section, are present in a given market; hence, it answers the question: how likely is it that collusion *will form*? In contrast, the behavioral approach answers the question: how likely is it that collusion *has formed*? In other words, it checks whether observed behavior is consistent with collusive behavior and whether there are competing theories that could also explain the observed patterns.

An industry for which there is past evidence of collusion, or even attempts to sustain collusion, should be more vulnerable to collusion in the future too. In this case, a merger would tend to facilitate collusion even more.

In order to check whether this is the case, behavioral collusive markers could prove useful.⁵¹ Collusive markers involve looking at data of certain variables, mainly prices and market shares to see whether their pattern is consistent with either tacit or explicit collusion.

Since the ultimate aim of colluding firms is to raise prices, unusually high prices might provide some hint of collusion. The problem is that it is not always possible to construct the correct *contrafactual*, i.e., the price that would have prevailed in a competitive environment. For this reason, one should compare industry prices with those of a control group with similar costs and characteristics. For instance, as reported in Abrantes-Metz and Bajari (2009), organized crime in New York created during the 1980s a “concrete club” that led to prices that were 70 percent higher than in other large cities: even taking into account the higher New York prices, the comparison suggested suspiciously high prices.

The fact that prices do not reflect costs might also be very informative. Indeed, theory suggests that in competitive environments prices tend to track costs of production. Bayari and Ye (2003) show that in a first-price sealed-bid auction with private values, the equilibrium bids are a function of costs when firms behave competitively. Instead, in an efficient cartel, firms would share their cost estimates, and then the lowest cost firm would submit a serious bid whereas all the other cartel members would either refrain from bidding or submit high “phony” bids.

Athey et al. (2004) analyze a model where firms’ costs are independent and identically distributed (i.i.d.) over time and are private information. Colluding firms would exchange messages over their costs before setting prices. They would then face a trade-off between efficiency (optimally, it is the lowest cost firm that would sell) and the price level: if the price is high, a high cost would declare a low cost; hence, truthful revelation would require choosing a low enough collusive price, but this mechanism might be too costly in terms of foregone profits. The authors show that at the best collusive equilibrium, provided that firms are patient enough, collusion entails stable prices and market shares over time.⁵²

These theoretical works suggest therefore that if prices do not track costs, there might be collusion in the industry. This explains why, for instance, an antitrust authority might want to look at the evolution of prices and costs over time. For instance, in the *DS Smith/Linpac Containers* merger case (2004), the UK Competition Commission looked at the time series of

⁵¹ On collusive markers (or screens) see Harrington (2006b) and for a less informal discussion Abrantes-Metz and Bajari (2009).

⁵² At a more general and intuitive level, one could say that price rigidity can also reflect the fact that agreeing to adapt to changing market conditions is difficult and costly (e.g., communication leaves traces that authorities can use to detect cartels).

DS Smith's unit prices and costs – and since changes in prices followed changes in costs quite closely, it concluded that it did not offer evidence of collusion (buyers claimed that there was collusion in the industry).

Related to the above-mentioned theoretical results that collusion would involve a higher price stability, Abrantes-Metz et al. (2006) have developed a *variance screen* of collusion. The analysis of a cartel in procurement auctions for food supply to military agencies in the USA, revealed that prices in frozen perch were much less volatile (and less responsive to costs) during the life of the cartel than when the cartel broke down.

At the other extreme, abrupt increases in prices that are not justified by cost or demand shocks may indicate that the industry is colluding. However, as Harrington (2006b) warns, cartels are aware that unusual price changes would attract unwanted attention, and accordingly often adopt progressive price increase policies.

Similarly, abrupt price decreases might also denounce the presence of a cartel. The occurrence of price wars (i.e., periods of intense rivalry followed by the return to a stable path of higher prices) as explained in Section 2.2.3, is a necessary component of collusion in markets in which transparency is low: price wars are used as a disciplining device to avoid deviations.⁵³ Price wars could also be indicative of failed attempts to collude. In contrast, the absence of price wars should not be considered as conclusive evidence of competitive behavior, given that price wars are costly and the most successful cartels are characterized by price stability.

Collusive price patterns also translate into distinctive *output patterns*. Indeed, quantity markers shed light on whether collusion took place or not by looking at the evolution of market shares. Under collusion, firms' market shares tend to be stable.⁵⁴ Also, the birth and the death of a cartel might give rise to abrupt changes in market shares and thus be indicative of a change in behavior from competition to collusion or viceversa.

It is important to notice that evidence consistent with collusion does not *prove* that collusion indeed took place, and the analysis should be careful enough to exclude any alternative plausible explanation of the observed behavior. Indeed, a sudden price reduction may not be due to the triggering of a price war in a Green and Porter-like cartel (Green and Porter, 2010), but may be due to a demand or cost shock. For instance, in the *Woodpulp* case (1988), it turned out that the alternating phases of high and low prices were caused by exogenous events such as shocks in the North American market that affected imports to Europe, and Swedish changes in the policy of subsidizing stocks (see Motta, 2004).

In any case, we should bear in mind that in a coordinated effects case, the purpose is not to prove that a cartel was in place, but rather that there is a sufficiently high probability that the merger is creating or strengthening collusion. Therefore, price and market share data that are consistent with collusive behavior should be taken as very serious evidence that collusion is likely to already exist in the industry.

⁵³ See Porter (1983) and Ellison (1994) for seminal empirical analysis of price wars and collusion in the Joint Executive Committee that operated in the USA at the end of the nineteenth century. Fabra and Toro (2005) empirically analyze price wars in the Spanish electricity market and show that they are consistent with collusion among electricity producers.

⁵⁴ If the market under scrutiny is a procurement auction, bid rotation might appear at first sight as resulting in negative correlation in firms' output levels. However, bid rotation would typically be constructed so as to guarantee stable market shares overall.

3.3 Other Approaches

Unfortunately, there have been few attempts to develop practical tools to measure the magnitude of coordinated effects.⁵⁵ The state of economic analysis in this area is still limited, and there is no consensus yet on how this issue should be approached from a quantitative perspective. However, for completeness, we report here three attempts to contribute to the measurement of coordinated effects.

3.3.1 Coordinated price pressure index

Price leadership is one way that firms can achieve coordination without explicit communication. In other words, as has been reported in some cases, one firm takes the lead in raising prices and the other firms match the price increase; failure to do so implies reversion to competitive pricing. Still, firms have to solve a coordination problem: namely, who will be the leader (Lu and Wright, 2010 and Harrington, 2012a).

Accordingly, it might be useful to quantify the incentives for a firm to take the lead in initiating collusion and how a merger impacts on such incentives. This is the approach followed by Moresi et al. (2011), who develop an index – referred to as the *coordinated price pressure index* (CPPI) – that is the largest price increase that a firm would be willing to initiate and its rival would be willing to match. A high CPPI indicates high chances that firms achieve collusive outcomes through price leadership.

In merger analysis, one would need to compute the delta CPPI, which is the increase in the CPPI that results from a merger. If the CPPI significantly increases from the pre- to the post-merger market structure, the merger can be expected to lead to coordinated effects.

For the sake of simplicity, the construction of the CPPI rests on strong assumptions. For instance, it does not look at the incentives to initiate a price increase in a fully dynamic model among all the firms in the industry, but instead focuses on two firms' incentives to raise and match the price increase in a single round. If firms are asymmetric, the CPPI can differ depending on the identity of the leader, and caution leads to taking the lowest value of the resulting CPPI.

The data needed to compute the CPPI include sales volumes, own price elasticities, diversion ratios,⁵⁶ profit margins and the discount factor.⁵⁷ Moresi et al. (2011) provide the exact formula to compute the CPPI, as well as several examples that illustrate how it can be computed.⁵⁸ For instance, consider two firms that compete by choosing prices; they have equal sales, and charge a margin of 40 percent. Their products are such that the diversion ratio between them is 25 percent, and the discount factor is 80 percent. The maximum price increase that each firm is willing to undertake is 10 percent, while the highest price increase that each firm is willing to match is 10.7 percent. Hence, the CPPI is 10 percent. Suppose that one of these two firms proposes to merge with another one. If the CPPI increases to 15 percent, then the delta CPPI would be 5 percent; in other words, the merger would facilitate collusion

⁵⁵ This is in contrast with the analysis of unilateral effects, for which a number of simple tests now exist to assess the effect of a merger on the pricing behavior of the merging firms. See Oxera (2011) for a review of such tools.

⁵⁶ The diversion ratio measures how much of the displaced demand for product A switches to product B when the price of A goes up.

⁵⁷ These ingredients are also used to compute indexes for the assessment of unilateral effects in merger cases, e.g., the gross upward pricing pressure index (GUPPI).

⁵⁸ They also apply to the attempted 2011 merger between AT&T and T-Mobile.

through price leadership by increasing by 5 percent the maximum price increase that firms are willing to lead and to match.

It is not fully clear whether the CPPI measures the likelihood of coordinated effects, or their magnitude conditional on such effects being likely. As for the likelihood, one would like to know the effects on profitability, which need not coincide with the magnitude of the price increases.⁵⁹

3.3.2 Incremental payoffs from collusion

Aubert et al. (2006) advocate for an alternative analysis. They argue that quantifying the incremental payoffs from post-merger collusion among subsets of firms in the post-merger market would provide valuable information as to whether coordinated effects are more or less likely. This is grounded in the assumption that the probability of coordination will be greater the higher the payoff from doing so; but otherwise, their analysis does not require a direct quantification of the likelihood of post-merger coordination. Their approach requires selecting a model of competition, and calibrating it using pre-merger data.

They provide an example of a market in which two out of four firms decide to merge. Under the assumption of differentiated products price competition, they compute equilibrium profits pre-merger, post-merger under no collusion, and post-merger under collusion among different subsets of firms. They find, under a specific parametrization, that collusion after the merger would be more than three times more profitable than collusion before the merger. Evidence showing that the payoffs from incremental collusion increase substantially after the merger, would indicate a strong likelihood of coordinated effects.

3.3.3 Diversion ratios and cross-price elasticities

Ivaldi and Lagos (2017) rely on numerical simulations to obtain predictions regarding the assessment of coordinated effects in merger cases. In particular, the authors simulate 50,000 markets, with 10,000 consumers and five single-brand firms in each, under the assumption that consumer preferences behave according to a model of discrete choice demand with random coefficients. On the basis of these simulations, the authors identify the factors enhancing the coordinated effects and the metrics that would allow for a better screening of mergers.

Their focus is on the effects of mergers on the critical discount factor above which a firm would find it optimal not to deviate from a trigger strategy sustaining perfect collusion. The paper proposes to decompose the impact as the sum of two effects: the change in profits (CP) and the asymmetry in payoffs (AP) effects. The former captures the change in the critical discount factor due to merging firms' internalization of the price effects on the merging brands; the latter captures firms' asymmetries as reflected in their different critical discount factors prior to the merger. Which of the two effects dominates depends on the symmetries/asymmetries among the merging firms. For symmetric mergers, the change in the critical discount factor post-merger is given by the CP effect alone, given that the AP effect is zero (pre-merger, all firms have the same discount factor). In contrast, when the merging firms are fairly asymmetric, the AP is the dominant effect because the internalization effect

⁵⁹ In line with this, Ivaldi and Lagos (2016) argue that the CPPI does a poor job in predicting coordinated effects. They show that it is important to incorporate information about the diversion ratios among the products of the merging firms, as they affect the costs of initiating the price increase after the merger. In particular, a merger between firms with high diversion ratios is more likely to make a price increase profitable. Because the CPPI does not incorporate this information, they claim that it fails to predict coordinated effects.

is weaker. The paper confirms that mergers involving symmetric firms, with high diversion ratios among their products, are likely to be worrisome. It also raises concerns about mergers between asymmetric firms with high cross-price elasticities, particularly so when one of the merging firms is a maverick (i.e., a small firm who would otherwise have disrupted collusion).

These differences across symmetric versus asymmetric mergers have implications for the types of indexes that are better at capturing the likelihood of coordinated effects. The authors show that, in the case of mergers among symmetric firms, the diversion ratios pre-merger are a good proxy as they capture the internalization effect across the merging brands. In contrast, in the case of asymmetric mergers, the cross-price elasticities are a good predictor of coordinated effects, and it is superior to using the merging firms' market shares.

4 COORDINATED EFFECTS IN EUROPEAN MERGER POLICY

In this section, we briefly report on the use of coordinated effects in European merger control, and conclude with a description of the *ABF/GBI* merger case.

4.1 The European Merger Regulation of 1989

4.1.1 The dominance test

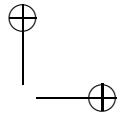
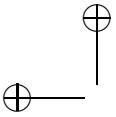
When merger control was finally introduced in the late 1980s, the criterion to authorize or prohibit mergers in Europe was based on the concept of *dominance*, that is, “the power to behave to an appreciable extent independently of its competitors, its customers and ultimately of the consumers.”⁶⁰ (In practice, for a finding of dominance a firm must enjoy a very high degree of market power, and it is widely accepted that it is unlikely that a firm with less than 40 percent of the relevant market would be found dominant.)

The Merger Regulation used to state that “a concentration which creates or strengthens a dominant position as a result of which effective competition would be significantly impeded in the Common Market or in a substantial part of it shall be declared incompatible with the Common Market.”⁶¹ In other words, only mergers that created or reinforced a dominant position could be prohibited by the European Commission. This introduced a test that is different from the “substantial lessening of competition” test used in US law and more aligned with economic analysis.

To see why the two tests may well lead to different outcomes when applied to the same merger, consider a situation where two or more firms with sizeable market shares would coexist in an industry after a merger, but none of them has enough market power to be considered dominant, and suppose it is also very unlikely that they would collude. For

⁶⁰ See *Hoffmann-La Roche* (1979), where the European Court of Justice first defined this concept thus: “The dominant position ... relates to a position of economic strength enjoyed by an undertaking which enables it to prevent effective competition being maintained on the relevant market by affording it the power to behave to an appreciable extent independently of its competitors, its customers and ultimately of the consumers. Such a position does not preclude some competition, which it does where there is a monopoly or quasi-monopoly but enables the undertaking which profits by it, if not to determine, at least to have an appreciable influence on the conditions under which that competition will develop, and in any case to act largely in disregard of it so long as such conduct does not operate to its detriment.”

⁶¹ Merger Regulation 4064/89, Article 2(3). Note that the legal term “concentration” stands for merger (or takeover).



instance, imagine that a firm has 50 percent, the two merging firms would have a share of 45 percent after the merger, while the remaining market is fragmented among smaller firms. In such a situation, economic theory clearly indicates that the merger might well be detrimental because of unilateral effects (suppose, for instance, that the enhanced market power is not outweighed by efficiency gains), but it would be very hard to argue that the merger would create or reinforce a dominant position, since the merging firms would face a stronger competitor. Hence, the Commission could not prohibit such a merger, as under the Merger Regulation 4064/89 the finding of a dominant position was a necessary condition for prohibiting a merger.

4.1.2 Joint dominance

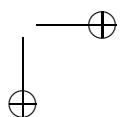
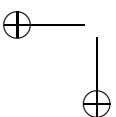
Soon, the European Commission realized that there were mergers that did not appear to be “good” (because they reduced competition, and were likely to raise prices) but that could not be prohibited because they did not create or reinforce a *single-firm*’s dominant position. However, the Commission could still prohibit such a merger if it could argue that it created or reinforced a *joint* dominant position. Loosely speaking, joint dominance refers to a situation where a (presumably small) group of firms in the market are able to coordinate their actions and set prices above the competitive level. However, what exactly joint dominance was, and how it could be proved to exist (or to likely occur after a merger), became the object of a series of merger cases in the EU.

The first case where the Commission challenged a merger on joint dominance grounds was *Nestlé/Perrier* (1992), a merger in the French mineral water industry. This was a case where all the elements pointed to high likelihood of coordination (probably pre-existing the merger) among the main firms, but the Commission eventually allowed the merger under some remedies, probably with a view to establishing a precedent that would not be challenged in court. After *Nestlé/Perrier* it was uncertain for a while whether the Community courts would uphold the Commission’s argument that a merger may be prohibited because of *joint* dominance.

In *France v. Commission* – a 1998 judgment – the European Court of Justice accepted the concept of joint dominance, but then quashed the decision that the Commission had taken in *Kali+Salz/MdK/Treuhand* (1999) and seemed to indicate that some sort of structural links (“correlative factors”) among firms was needed to prove joint dominance. Although it was unclear what exactly and how strong such structural links should be, this judgment seemed to set a very high standard to prove joint dominance.

However, in *Gencor v. Commission* (1999) the Court of First Instance (CFI) reaffirmed the principle that the European Commission can block mergers if they create joint dominance but seemed to accept a broader (and more economics-aligned) interpretation of the concept, and argued that there is no need for oligopolists to have some structural links in order to prove that collective dominance exists.

The Court stated that “the concentration would have had the direct and immediate effect of creating the condition in which abuses were not only possible but economically rational, given that the concentration would have significantly impeded effective competition in the market by giving rise to a lasting alteration to the structure of the markets concerned” (para. 94 of Judgment). The judgment seemed to pay less attention to structural links between the firms and more attention to the structure of the market, referring in particular to the fact that the merger would have rendered the position between the two main producers extremely



symmetric, both in terms of reserves of world platinum production and in terms of costs of production.

The Commission was then ready to use the higher degree of freedom left by the CFI judgment, and started to increasingly rely on the concept of joint dominance, applying it to cases where it was not straightforward that the merger would have created or reinforced collusion. Arguably, though, joint dominance was the only tool available to the Commission to prohibit anticompetitive mergers that it could have not otherwise stopped.

4.1.3 Airtours, and the new regulation

The *Airtours* judgment of the Court of First Instance (2002) (followed immediately by other two judgments, *Schneider/Legrand* and *Tetralaval/Sidel*, 2003, in which the CFI also annulled merger prohibition decisions of the Commission) is very important because it led to a change in EU merger policy.

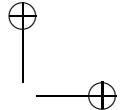
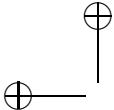
In *Airtours*, the Commission had extended the concept of joint dominance to an industry whose features were not unambiguously conducive to collusion. The CFI went very carefully through the economic analysis of the Commission, and annulled the decision. Its judgment contains a number of remarkable points.

First, the CFI clarifies the standards of proof required by a merger prohibition: it is not enough for the Commission to argue that after the merger it is *possible* that firms will collude, it should motivate and explain that the collusive outcome will be very *likely* to arise. (Similarly, in *Tetralaval/Sidel* – where, however, the issue was whether the merger would have led to anticompetitive tying – the Court stressed that the standard of proof cannot consist in showing the *mere possibility* that a certain outcome can occur, but requires strong arguments and evidence that such an outcome would be *plausible*.)

Second, this judgment makes it clear that joint dominance is not a multi-purpose concept, but has to do with the pro-collusive effects of a merger, as economic analysis would have shown. In particular, the Court spells out three conditions for tacit coordination to be sustainable: (i) sufficient market transparency (for firms to monitor each other and see whether there are deviations); (ii) the existence of an incentive not to depart from the common policy, i.e., the existence of a credible mechanism of retaliation if deviations occur; (iii) current and prospective rivals, as well as consumers, must not jeopardize coordination (in other words, neither entry is easy nor buyer power very high). These are the same conditions that any economic textbook would indicate as those that allow for a collusive outcome to arise. Therefore, the judgment clarifies once and for all that the concept of joint dominance used by the European judges is the same as the one used in economic analysis.

Finally, in this and the following judgments the CFI heavily criticizes the economic analysis carried out by the Commission, persuading Commissioner Mario Monti that the use of economics and economists at DG-Competition should be enhanced, and to create the Chief Economist's Office.

After *Airtours*, it was clear that the Commission could not rely too much on the joint dominance concept to prohibit mergers that it regarded to raise anticompetitive concerns but that did not create or strengthen a single-firm dominant position. This pushed it to adopt a new Merger Regulation (entered into effects in May 2004) with a new test for the assessment of merger control: the Commission will prohibit mergers that “would significantly impede effective competition, in the common market or in a substantial part of it, in particular as a result of the creation or strengthening of a dominant position.”



116 *Handbook of game theory and industrial organization: applications*

In part not to lose the case law, in part to accommodate the objections of some member states (the dominance test still applies in some national laws), a reference to “dominance” is kept, but the “test” *de facto* is modified from a dominance test to a “substantial lessening of competition” test.

4.2 The Horizontal Merger Guidelines

The Horizontal Merger Guidelines (HMGs) issued in 2004 by the Commission follow the conditions for coordinated effects as set out by the CFI in *Airtours* and subsequently confirmed by the Court of Justice in *Impala* (2006), and that are to a large extent consistent with what economic analysis suggests (see Section 2.1 above):

Coordination is more likely to emerge in markets where it is relatively simple to reach a common understanding on the terms of coordination. In addition, three conditions are necessary for coordination to be sustainable. First, the coordinating firms must be able to monitor to a sufficient degree whether the terms of coordination are being adhered to. Second, discipline requires that there is some form of credible deterrent mechanism that can be activated if deviation is detected. Third, the reactions of outsiders, such as current and future competitors not participating in the coordination, as well as customers, should not be able to jeopardise the results expected from the coordination. (Para. 41)

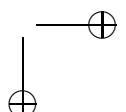
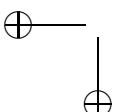
In other words, the Commission identifies the ability to reach some sort of common understanding (on prices, on capacities, on terms of sales, on how to divide markets, and so on) as a precondition for coordinated effects, followed by the three (cumulative) conditions for the sustainability of the collusion, namely (i) a mechanism or circumstances that allow monitoring of each other’s actions; (ii) the ability and credibility of a mechanism that allows the punishment of deviations; and (iii) the inability of customers to command lower prices and of existing or prospective rivals to react, thus making it unlikely to reach the collusive outcome.⁶²

The HMGs also clarify that the merger may raise coordinated effects concerns in two ways: (i) by increasing the likelihood that firms will (tacitly or explicitly) coordinate their behavior *after* the merger (e.g., because the merger reduces the number of existing competitors, increases the symmetry of the main firms aligning their incentives to collude, or removing a maverick firm that in the past had prevented or threatened collusion); or (ii) by making coordination that already existed before the mergers easier, more stable or more effective.

Finally, the HMGs also point out that efficiency gains could well have a procompetitive effect not only in unilateral effects but also in coordinated effects cases: “In the context of coordinated effects, efficiencies may increase the merged entity’s incentive to increase production and reduce prices, and thereby reduce its incentive to coordinate its market behavior with other firms in the market. Efficiencies may therefore lead to a lower risk of coordinated effects in the relevant market” (para. 82).

To the extent that the Commission follows the HMGs, and it applies the analysis not in a mechanical way (prior to *Airtours*, one could get the impression that most of the analysis had consisted in a listing of the main facilitating factors without really trying to uncover the real workings of the market and the degree to which collusive outcomes may be plausible and

⁶² In general, economic theory suggests that there are two important aspects of collusion, namely enforcement and coordination. In merger analysis, though, it is enforcement that should be the focus of the analysis, whereas in anticompetitive agreements and cartels (or conspiracies, in US law), the focus is on coordination.



sustainable), coordinated effects in EU competition policy will be aligned to the teachings of economic analysis. An indication that the Commission is starting to do a good job in this direction comes from the recent *ABF/GBI* case (2008), as described next.

4.3 ABF/GBI: Application of the 2004 European Guidelines

The *ABF/GBI* merger (2008) was the first merger challenged (but eventually approved subject to sizeable remedies) by the European Commission on the basis of coordinated effects since *Airtours*. In this case, the Commission had the chance to apply its own Merger Guidelines, which in turn were modeled after the *Airtours* judgment (later affirmed by the European Court of Justice).

It is interesting not only because it is illustrative of the way in which EU merger control is enforced, but also because it shows the importance of a careful analysis of the industry and how differences in some features of the market may lead to very different outcomes of the investigation (notably, differences in the distribution sector in Spain and Portugal relative to France led to an assessment of coordinated effects in the former but not in the latter).

The case concerned the acquisition of GBI's yeast⁶³ operations in Continental Europe⁶⁴ by Associated British Foods (ABF), and the Commission's investigated the merger upon referral from the Spanish, Portuguese and French authorities. Accordingly, the relevant markets were defined as those for compressed, dry and liquid yeast in each of these three countries. (We shall focus on compressed yeast, which is the most important.)

The Commission's assessment, following the case law and the HMGs, hinges on three steps. First, it analyzes the basic features of the market, to see whether they are conducive to coordination. Second, it studies whether any coordinated outcome would be sustainable. This step itself requires – in the light of *Airtours* – to show that (i) the market is sufficiently transparent to allow monitoring of deviations; (ii) there exists a credible mechanism to punish them; and (iii) it is unlikely that outsiders (be they customers or entrants) may prevent tacit or explicit collusion. Third, it must show that the merger either strengthens coordination (if it already exists) or makes it more likely. We discuss these three steps in the following subsections.

4.3.1 Market features make coordination likely

The decision mentions a number of features of the market that are likely to be conducive to coordinated behavior. There is a high degree of concentration, with ABF and GBI's combined market share being around 70–80 percent in Portugal, 40–50 percent in Spain and 30–40 percent in France, while French-based yeast group Lesaffre's shares were respectively 20–30 percent, 40–50 percent, and 60–70 percent. The market is also characterized by frequent interaction (in Spain and Portugal, buyers are mostly small artisanal bakers who cannot afford refrigerated storage and order yeast with a weekly or bi-weekly frequency); products are homogeneous (although in France Lesaffre seems to enjoy a higher-quality status); demand is stable or declining; it is unlikely that new technologies may break the market equilibrium; in Spain and Portugal (but not in France, where bakery is no longer

⁶³ Yeast is an essential ingredient in the production of bread and bakery products. It is perishable and even when refrigerated it lasts only for three to four weeks.

⁶⁴ GBI's yeast business in the UK and South America were sold to French-based yeast group Lesaffre, and approved (subject to remedies) in a prior merger investigation.

artisanal, and distribution is in the hands of centralized groups), there is small buyer power; there are barriers to entry and expansion (production has become increasingly concentrated in fewer plants, witnessing economies of scale); and multi-market contacts across Europe exist among all the main players.

The analysis of past price and output data also revealed significant market share stability and price parallelism even when production was hit by input cost increases. As the Commission puts it: “Such supply shocks can, in some circumstances, disrupt any efforts to tacitly coordinate conduct, particularly to the extent that they may affect some players more than others. However . . . given common technology and climatic conditions of the plants of ABF, GBI and Lesaffre serving the Spanish market, increased input costs can be expected to affect all three players in a similar manner” (para. 224). Interestingly, but not surprisingly, internal documents revealed that firms were fully aware of their symmetry in this respect, and that therefore their interests in price increases were perfectly aligned.

4.3.2 Sustainability of coordination

Although the “checklist” of the factors that may facilitate collusion is a useful step in the investigation, the crucial step is then to understand how likely it would be that deviations may be monitored and punished:

1. In this case, the distribution sector plays a fundamental role in determining the degree of *transparency* of the market. The Commission found that the Spanish and Portuguese markets were characterized by very strong and stable relationships both between distributors and their clients (in many cases, very strong personal relationships developed over time, due also to the frequent visits of the distributors) and between producers and distributors (which were *de facto* or *de jure* exclusive dealers and which enjoyed exclusive territorial protection), the latter also reporting information on market developments to the former (often, reporting information back to the suppliers was part of the distribution agreement or was incentivized in contracts).⁶⁵ On the other hand, distribution in France was in the hands of concentrated and centralized groups that bought from several suppliers and served industrial buyers. The Commission stressed how the simple organization of the distribution sector in Spain and Portugal allowed suppliers to efficiently monitor the market,⁶⁶ whereas in France such transparency could not be achieved.
2. As for the capacity to deter deviations through *credible punishments*, the Commission found that “all three players – GBI, ABF and Lesaffre – currently hold excess capacity in their plants serving Spain, sufficient to initiate a long-lasting price war in the event of any of them deviating from coordinated interaction” (para. 242). If necessary, they could have

⁶⁵ Although bakers had a primary distributor/supplier, they also developed some relations with, and minor purchases from, a secondary source. Yeast being indispensable, this was a way for bakers to ensure themselves against possible shortages or failures in primary sourcing. In turn, this link with another distributor/supplier allowed bakers to switch supplier in case the primary increased prices. But in turn, this would mean that the primary distributor/supplier may be informed of possible “deviations” by rivals.

⁶⁶ The role of a stable demand in increasing transparency of the market is clearly explained in the following excerpt from the Decision: “In the context of frequent deliveries, [monitoring deviations] is simply verified by observing significant decrease in volumes with respect to the previous year for a given territory. Indeed when market demand is relatively stable, as is the case in Spain, inferring deviations from collusive conduct is easier and requires less market data than when the market demand fluctuates significantly and unpredictably” (para. 232).

also used capacity in plants located elsewhere.⁶⁷ Furthermore, retaliation would have been timely given that the high frequency of market transactions, and its threat would have been enhanced by the existence of multi-market contacts.

3. As for the *reactions of outsiders*, the third of the conditions stressed by *Airtours*, the Commission found that the (fragmented) competitors as well as importers were facing high barriers to entry and expansion; and that there was limited countervailing buyer power of distributors (that as we have seen were linked by exclusive deals to producers); and bakers (who were mostly small artisans).

4.3.3 Coordinated effects of the merger

Last, “the Commission must further show, on the basis of a prospective analysis, the extent to which the ‘alteration in the [relevant market] structure that the transaction would entail’ [*Airtours*, para. 61] significantly impedes effective competition by making coordination easier, more stable or more effective for the three firms concerned either by making the coordination more robust or by permitting firms to coordinate on even higher prices” (para. 273). In this respect, it found the following:

1. The merger increased transparency by reducing the number of players, facilitating the detection of deviations and retaliations (when only two firms exists, there is no risk of free-riding in the punishment efforts, nor possibility to make mistakes on the identity of the deviators).
2. GBI exhibited differences relative to ABF and Lesaffre. First, GBI served Spain and Portugal from its Italian plant, which also served other markets. This means that demand and supply shocks affecting other markets may have repercussions on the Iberian markets. After the merger, ABF/GBI would reorganize production relying on local plants, thereby removing this possible source of misalignments facing shocks.

Second, it had made a number of improvements in production and packaging. However, under the terms of the merger agreements, GBI’s patents would be shared by ABF and Lesaffre, which by doing so “(a) eliminate GBI as a source of potentially destabilizing innovation and (b) ensure neither of the two coordinating firms inherits the competitive advantage that may eventually derive from IP rights” (para. 301).

Third, GBI was not present in the market for liquid yeast, mostly used to supply industrial bakers. In case of growth of this market relative to compressed yeast, this may have been a further source of misalignment of incentives.

In general, after the merger ABF/GBI and Lesaffre would be highly symmetric in terms of production costs, capacities,⁶⁸ and market shares (both of them would have 40–50 percent), thereby facilitating tacit collusive outcomes.⁶⁹

4.3.4 The remedies

On the basis of the above-mentioned analysis, the Commission concluded that the merger would have created or strengthened coordinated effects in Spain and Portugal (but not, as

⁶⁷ “Shifting volumes from one geographic market to the other, though likely uneconomical on a permanent basis given the opportunity cost of lost sales, allows the three producers to reinforce the threat of significantly expanding sales without necessarily holding excessive idle capacity” (para. 242).

⁶⁸ After the merger “both Lesaffre and ABF would have almost identical spare capacities … in the Iberian Peninsula” (para. 297).

⁶⁹ Symmetry would instead be absent post-merger from the French market, largely dominated by Lesaffre.

we saw, in France). Still, the transaction was cleared subject to the remedies proposed by the parties. An initial remedy consisted in the divestment of GBI's sales and distribution activities in Spain and Portugal, but did not include a production plant (it only included an agreement to supply the buyer for three years with yeast produced at GBI's Italian plant), but it was not accepted because the lack of a production plant would have not made the buyer a serious competitor. Ultimately, the accepted remedy consisted in offering, on top of sales and distribution assets, either a UK plant or the plant located in Portugal.⁷⁰

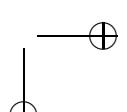
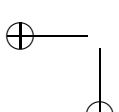
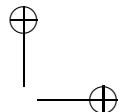
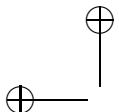
5 CONCLUSIONS

Mergers lead to coordinated effects when they increase the likelihood that firms will reach (tacit or explicit) collusive outcomes in the post-merger market. Therefore, a careful assessment of coordinated effects is necessary in order to prevent anticompetitive mergers from taking place.

We have reviewed the main factors that, from an economic point of view, should be analyzed in a coordinated effects analysis. The main questions to be addressed are whether collusion would be sustainable after the merger, and how the merger contributes to the sustainability of collusion. Certain supply factors – such as a small number of symmetric firms, barriers to entry, or multi-market contact – and demand factors – such as demand stability and the existence of regular and frequent orders – contribute to facilitating collusion. Also, price transparency on the sellers' side and communication about past and future conduct make it easier for firms to reach and respect a collusive agreement. A merger that takes place in a market already conducive to collusion, is likely to enhance collusion and thus raise concerns over coordinated effects. The incidence of some mergers on the likelihood of collusion might be stronger than others: particularly worrisome are those that increase symmetry in markets in which there are already few competitors. The assessment of coordinated effects in vertical merger cases points out that vertical integration should raise more concerns when it involves relatively large buyers in markets in which producers have little information regarding retail markets. From an applied perspective, the quantification of coordinated effects in merger cases is an area in economics that is not yet fully developed, and while some simple indexes now exist, there is no unanimity about their usefulness.

The description of the European merger policy provides useful hints that can guide the assessment of coordinated effects in future mergers cases. The experience highlights the importance of identifying those mechanisms available to firms for monitoring compliance and to credibly punishing deviators in order to make collusion sustainable. It also stresses that the assessment of coordinated effects requires detailed knowledge of the industry, as well as a careful analysis of the past performance and interaction among firms in the market. To be sure, because the analysis is often very delicate and complex, reliance on economic theory should help us on correctly assess the likelihood of coordinated effects in merger cases.

⁷⁰ The latter turned out to be implemented: Lallemand, a German competitor with limited presence in Spain and Portugal, bought GBI's sales and distribution business as well as the Portuguese plant (see Neven and De la Mano, 2009).

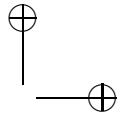
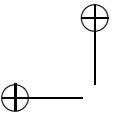


REFERENCES

- Abrantes-Metz, R. and P. Bajari. 2009. "Screens for Conspiracies and their Multiple Applications." *Antitrust*, 24: 66–71.
- Abrantes-Metz, R., L.M. Froeb, J.F. Geweke and C.T. Taylor. 2006. "A Variance Screen for Collusion." *International Journal of Industrial Organization*, 24: 467–486.
- Athey, S. and K. Bagwell. 2001. "Optimal Collusion with Private Information." *RAND Journal of Economics*, 32: 428–465.
- Athey, S., K. Bagwell and C. Sanchirico 2004. "Collusion and Price Rigidity." *Review of Economic Studies*, 71(2): 317–349.
- Aubert, C., P. Rey and W. Kovacic. 2006. "The Impact of Leniency and Whistleblowing Program on Cartels." *International Journal of Industrial Organization*, 24: 1241–1266.
- Bajari, P. and L. Ye. 1993. "Deciding Between Competition and Collusion." *The Review of Economics and Statistics*, 85(4): 971–989.
- Bernheim, B.D. and M.D. Whinston. 1990. "Multimarket Contact and Collusive Behavior." *RAND Journal of Economics*, 21: 1–26.
- Buccigrossi, P. and G. Spagnolo. 2007. "Corporate Governance and Collusive Behaviour." In W.D. Collins (ed.). *Issues in Competition Law and Policy*, Chicago, IL: American Bar Association, Antitrust Section.
- Coate, M. "Empirical Analysis of Merger Enforcement Under the 1992 Merger Guidelines." *Review of Industrial Organization*, 27(4): 279–301.
- Compte, O. and P. Jehiel. 2010. "The Coalitional Nash Bargaining Solution." *Econometrica*, 78(5): 1593–1623.
- Compte, O., F. Jenny and P. Rey. 2002. "Capacity Constraints, Mergers and Collusion." *European Economic Review*, 46: 1–29.
- Cooper, R., D.V. DeJong, R. Forsythe and T.W. Ross. 1992. "Communication in Coordination Games." *Quarterly Journal of Economics*, 107: 739–771.
- Doyle, M. and C. Snyder. 1999. "Information Sharing and Competition in the Motor Vehicle Industry." *Journal of Political Economy*, 107(6): 1326–1364.
- Ellison, G. 1994. "Theories of Cartel Stability and the Joint Executive Cartel." *RAND Journal of Economics*, 25: 37–57.
- Engel, C. 2007. "How Much Collusion? A Meta-analysis of Oligopoly Experiments." *Journal of Competition Law and Economics*, 3: 491–549.
- European Commission. 2004. "Guidelines on the Assessment of Horizontal Mergers under the Council Regulation on the Control of Concentrations between Undertakings." *Official Journal of the European Union*.
- European Commission. 2008. "Guidelines on the Assessment of Non-horizontal Mergers under the Council Regulation on the Control of Concentrations between Undertakings." *Official Journal of the European Union*.
- Evans, W.N. and I.N. Kessides. 1994. "Living by the 'Golden Rule': Multimarket Contact in the U.S. Airline Industry." *Quarterly Journal of Economics*, 109: 341–366.
- Fabra, N. 2005. "Collusion with Capacity Constraints Over the Business Cycle." *International Journal of Industrial Organisation*, 4(1): 69–81.
- Fabra, N. and J. Toro. 2005. "Price Wars and Collusion in the Spanish Electricity Market." *International Journal of Industrial Organisation*, 23(3–4): 155–181.
- Farrell, J. and C. Shapiro. 1990. "Horizontal Mergers: An Equilibrium Analysis." *American Economic Review*, 80(1): 107–126.
- Fonseca, M.A. and H.-T. Normann. 2008. "Mergers, Asymmetries and Collusion: Experimental Evidence." *The Economic Journal*, 118: 387–400.
- Gilo, D., Y. Moshe and Y. Spiegel. 2006. "Partial Cross Ownership and Tacit Collusion." *RAND Journal of Economics*, 37(1): 81–99.
- Green, J. and R.H. Porter. 2010. "Noncooperative Collusion under Imperfect Price Information." *Econometrica*, 52(1): 87–100.
- Haltiwanger, J. and J.E., Jr. Harrington. 1991. "The Impact of Cyclical Demand Movements on Collusive Behavior." *RAND Journal of Economics*, 22: 89–106.
- Harrington J.E., Jr. 2005. "The Collusion Chasm: Reducing the Gap Between Antitrust Practice and Industrial Organizational Theory." *CSEF-IGER Symposium on Economics and Institutions*.
- Harrington, J.E. 2006b. "Behavioural Screening and the Detection of Cartels." *European University Institute, 2006 EU Competition Law and Policy Workshop/Proceedings*.
- Harrington, J.E., Jr. 2012a. "Evaluating Mergers for Coordinated Effects and the Role of 'Parallel Accommodating Conduct'." *Johns Hopkins University Working Paper No. 601*.
- Harrington, J.E., Jr. 2012b. "A Theory of Tacit Collusion." *Johns Hopkins University Working Paper No. 588*.
- Harsanyi, J.C. and R. Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

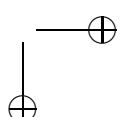
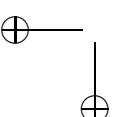
122 *Handbook of game theory and industrial organization: applications*

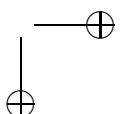
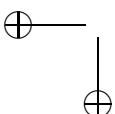
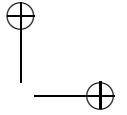
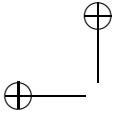
- Huck, S., H.-T. Normann and J. Oechssler. 2004. "Two Are Few and Four Are Many: Number Effects in Experimental Oligopolies." *Journal of Economic Behavior & Organization*, 53: 435–446.
- Ivaldi, M. and V. Lagos 2017. "Assessment of Post-merger Coordinated Effects: Characterization by Simulations." *International Journal of Industrial Organization*, 53: 267–305.
- Jullien, B. and P. Rey. 2007. "Resale Price Maintenance and Collusion." *RAND Journal of Economics*, 38(4): 983–1001.
- Kaplow, L. and C. Shapiro. 2007. "Antitrust." In M.A. Polinsky and S. Shavell (eds), *Handbook of Law and Economics, Volume 2*, Amsterdam: Elsevier.
- Kühn, K.-U. 2004. "Coordinated Effects of Mergers in Differentiated Products Markets." *CEPR Discussion Paper No. 4769*.
- Kühn, K.-U. 2008. "On the Coordinated Effects of Conglomerate Mergers." Mimeo.
- Lepore, J. and C. Knittel. 2010. "Tacit Collusion in the Presence of Cyclical Demand and Endogenous Capacity Levels." *International Journal of Industrial Organization*, 28(2): 131–144.
- Levenstein, M.C. and V.Y. Suslow. 2002. "What Determines Cartel Success?" *Journal of Economic Literature*, 44: 43–95.
- Lu, Y. and J. Wright. 2010. "Tacit Collusion with Price-matching punishments." *International Journal of Industrial Organization*, 28: 298–306.
- Miklos-Thal, J. 2009. "Optimal Collusion under Cost Asymmetry." *Economic Theory*, 46: 99–125.
- Mezzanotte, F. 2009. "Can the Commission Use Article 82EC to Combat Tacit Collusion?" *East Anglia Working Paper 09-5*.
- Montero, J.P. and E. Johnson. 2012. "Multimarket Contact, Bundling and Collusive Behavior." *Pontificia Universidad Católica de Chile Working Paper No. 420*.
- Moresi, S.X., D. Reitman, S.C. Salop and Y. Sarafidis. 2011. "Gauging Parallel Accommodating Conduct Concerns with the CPPI." Washington, DC: Charles Rivers Associates.
- Motta, M. 2004. *Competition Policy. Theory and Practice*. Cambridge, UK: Cambridge University Press.
- Motta, M. and M. Polo. 2002. "Leniency Programs and Cartel Prosecution." *International Journal of Industrial Organization*, 21: 347–379.
- Motta, M. and Tarantino, E. 2016. "The Effect of a Merger on Investments." *CEPR Discussion Paper No. 11550*.
- Neven, D. and M. De la Mano 2009. "Economics at DG Competition, 2009–2010." *Review of Industrial Organization*, 37(4): 309–333.
- Nocke, V. and L. White 2007. "Do Vertical Mergers Facilitate Upstream Collusion?" *American Economic Review*, 97(4): 1321–1339.
- Nocke, V. and L. White 2010. "Vertical Merger, Collusion, and Disruptive Buyers." *International Journal of Industrial Organization*, 28: 350–354.
- Normann, H.-T. 2009. "Vertical Integration, Raising Rivals' Costs and Upstream Collusion." *European Economic Review*, 53: 461–480.
- Oxera. 2011. "Unilateral Effects Analysis and Market Definition: Substitutes in Merger Cases?" *Agenda: Advancing in Economics and Business*, July.
- Phillips, O.R. and C.F. Mason. 1992. "Mutual Forbearance in Experimental Conglomerate Markets." *RAND Journal of Economics*, 23(3): 395–414.
- Phillips O.R., C.F. Mason, and C. Nowell. 1992. "Duopoly Behavior in Asymmetric Markets: An Experimental Evaluation." *Review of Economics & Statistics*, 74(4): 662–700.
- Porter, R. 1983. "A Study of Cartel Stability: The Joint Executive Committee, 1880–1886." *Bell Journal of Economics*, 14: 301–314.
- Rotemberg, J.J. and G. Saloner. 1986. "A Supergame-Theoretic Model of Price Wars During Booms." *American Economic Review*, 76: 390–407.
- Rubinsten, A. 1982. "Perfect Equilibrium in a Bargaining Model." *Econometrica*, 50: 97–100.
- Snyder, C. 1996. "Negotiation and Renegotiation of Optimal Financial Contracts under the Threat of Predation." *Journal of Industrial Economics*, 44(3): 325–343.
- US Department of Justice and Federal Trade Commission. 1984. "Non-Horizontal Merger Guidelines."
- US Department of Justice. 1994. "Competitive Impact Statement: United States vs Airline Tariff Publishing Company et al." Civil Action No. 92-2854 (SSH).
- US Department of Justice and Federal Trade Commission. 2010. "Horizontal Merger Guidelines."
- Van Huyck, J.B., R.C. Battalio and R.O. Beil (1990) "Tacit Coordination Games, Strategic Uncertainty and Coordination Failure." *American Economic Review*, 80(1): 234–248.
- Vasconcelos, H. 2005. "Tacit Collusion, Cost Asymmetries and Mergers." *RAND Journal of Economics*, 36: 39–62.
- Vasconcelos, H. 2008. "Sustaining Collusion in Growing Markets." *Journal of Economics & Management Strategy*, 17(4): 973–1010.
- Whinston, M. 2006. *Lectures on Antitrust Economics (Cairolis Lectures)*. Cambridge, MA: MIT Press.

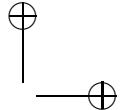
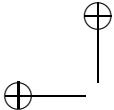


PART II

CONTESTS







6. Contest theory*

Luis C. Corchón and Marco Serena

1 INTRODUCTION

A contest is a game where contestants exert costly and irretrievable effort in order to obtain one or more prizes with some probability. Many real-life situations comply with this general definition. Political parties try to win the elections by spending resources on political campaigns. Plaintiffs and defendants try to win a trial by spending money on lawyers. Armies try to win wars by buying weapons and hiring soldiers. Lobbyists try to persuade policy makers by carefully preparing influential speeches. Students compete for scholarships by spending time studying hard. Sports people try to win sports competitions, or break world records, by tireless training. Job applicants try to get a job by giving their all on assessment day. And television quizzes, public procurements, R&D contests, and virtually countless other scenarios comply with the definition of a contest. In fact, contests can be regarded as an allocation mechanism on an equal footing with an authority or the market.

Contest literature traces back to the seminal contributions of Tullock (1967, 1980) and Krueger (1974) who studied a particular type of contest – rent-seeking – and of Becker (1983) who studied lobbying.¹ Complementary surveys over the literature on contests are: Nitzan (1994), Szymanski (2003), Corchón (2007) and Konrad (2009). We aim here to provide an up-to-date review of some of the main contributions in the field.

Contests are categorized into two big families. Contests that naturally occur in order to solve a dispute or a conflict – political campaign, court trial, war, lobby – and contests that are planned and organized by a contest designer in order to achieve some goal – scholarship, sport, job assessment day, television quiz, public procurement, R&D. In the latter family, the contest can to some extent be planned ahead by the designer who can choose, for instance, the number of participants. The optimal design of contests is considered in Section 7, whereas the rest of the chapter takes the structural elements of the contest as exogenous.

Formally, a contest is described by the following elements:

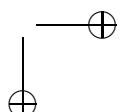
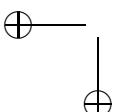
1. A list of contestants – or agents – denoted by $N = \{1, 2, \dots, n\}$.
2. The effort of each contestant, denoted by $G_i \in \mathbb{R}_+$ for contestant i .²
3. A prize whose value for contestant i is denoted by V_i .³ When all contestants have identical valuation of the prize it will be denoted by V .
4. A mapping from contestants' efforts into individual probabilities of winning the prize, or a share of it in the case that the prize is divisible. This mapping is called contest success

* We thank J. Atsu Amegashie, Subhasish Chowdhury, Matthias Dahm, Qiang Fu, Doron Klunover, Ron Siegel and Stergios Skaperdas for useful comments. Jana Bolvashenkova provided competent research assistance.

¹ A loose history of the concept of rent-seeking can be found in Tullock (2003).

² Effort could be multidimensional, as in Faria et al. (2014).

³ As for non-fixed prizes, see Chung (1996) and Amegashie (1999a). Prize valuation heterogeneities can be equivalently modeled as marginal cost heterogeneities, and they are often referred to as *types* of contestants.



function (CSF) and for contestant i it is denoted by $p_i = p_i(G_1, \dots, G_n)$. In Section 2 we discuss the three main CSFs, and their variations and alternatives. In Section 3 we discuss the microfoundation of CSFs.

5. An attitude towards risk of contestants. To the extend of keeping this survey simple and neat we will assume risk neutrality throughout.⁴
6. A cost function for effort, which for simplicity we assume here to be linear. Without loss of generality, we also assume that the marginal cost equals 1. The cost of effort is sunk.⁵

All in all, a contest can be represented as a normal-form game where players are contestants, strategies are efforts, and payoffs, denoted by Π_i , are the expected utility, that is:

$$\Pi_i(G_1, \dots, G_n) \equiv p_i(G_1, \dots, G_n) V_i - G_i \quad (6.1)$$

For these games the most common notion of equilibrium is the one proposed by John Nash in 1950, who generalized an idea of Antoine-Augustin Cournot in 1838: an equilibrium is a situation where there are no unilateral incentives to deviate.⁶ Hence, an n-tuple $(G_1^*, \dots, G_i^*, \dots, G_n^*)$ is a Nash equilibrium (NE) if

$$\Pi_i(G_1^*, \dots, G_i^*, \dots, G_n^*) \geq \Pi_i(G_1^*, \dots, G_i, \dots, G_n^*), \forall G_i \in \mathfrak{N}_+, \forall i \in N \quad (6.2)$$

Discussion of properties such as existence and uniqueness of NE and comparative statics can be found in Sections 3 and 4 of Córchón (2007).⁷ Sections 5 and 6 of the same survey investigate social welfare, and Section 7 analyzes the relation between rent-seeking, institutions and economic policies.

In the first half of this survey (Sections 2 and 3), we present the main ingredients of a contest, with a focus on how efforts translate into probabilities of winning (i.e., the CSF). In the last part of this survey, we focus on some extensions of the basic model, namely dynamics (Section 4), information (Section 5), and groups (Section 6), we compute the equilibrium of the popular lottery model of contest with heterogeneous contestants, and we use equilibrium properties to review some results on how the contest designer optimally designs the contest given her objective function, which depends on effort and possibly V_i s (Section 7).⁸

⁴ The interested reader can refer to Hillman and Katz (1984), Skaperdas (1991), Skaperdas and Gan (1995), Konrad and Schlesinger (1997), Cornes and Hartley (2003) and Treich (2010) for models of contest with risk-averse contestants.

⁵ A model in which expenses are partially reversible is to be found in Siegel (2010), where part of the effort is sunk and the rest is paid only by the winner of the contest. Despite our focus in this survey on all-pay contests, there is a complementary strand of the literature considering winner-pay contests; that is, the effort is exerted only if the contestant wins; see Skaperdas and Gan (1995), Wärneryd (2000), Córchón and Dahm (2011), Yates (2011) and Alcalde and Dahm (2013).

⁶ An alternative specification including bounded rationality of contestants is analyzed in a contest model by Anderson, Goeree and Holt (1998).

⁷ A recent work by Chowdhury and Sheremeta (2011) shows that uniqueness of NE is lost when there are externalities in costs.

⁸ The contest designer is typically assumed to design the contest *ex ante* with commitment. An exception is in Córchón and Dahm (2011), in which the CSF is derived as the optimal reply of a planner who cannot commit to the CSF.

2 CONTEST SUCCESS FUNCTION

2.1 Standard

In this subsection we present the three most common CSFs. An overview of CSFs' applications to econometric models can be found in Jia, Skaperdas and Vaidya (2013).

2.1.1 All-pay auction (Hillman and Riley, 1989)

In this version of the CSF the contestant exerting the highest effort wins the prize with probability 1. If several contestants exert the highest effort, it is usually assumed that they have an equal probability of winning the prize. Formally,

$$p_i(G_1, \dots, G_i, \dots, G_n) = \begin{cases} 1 & \text{if } G_i > \max\{G_1, \dots, G_{i-1}, G_{i+1}, \dots, G_n\} \\ \frac{1}{m} & \text{if } G_i \text{ is one of the } m \text{ maximum elements of } \{G_1, \dots, G_n\} \\ 0 & \text{if } G_i < \max\{G_1, \dots, G_n\} \end{cases} \quad (6.3)$$

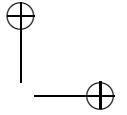
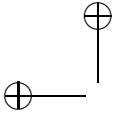
In all-pay auctions (henceforth, APAs) there are no equilibria in pure strategies. The reason is that for a given vector of efforts contestants have an incentive to decrease their efforts as long as it does not affect their winning probabilities, i.e., all the way down to 0 if the contestant's effort was not the highest one, or to an arbitrarily small epsilon more than the second highest effort if the contestant's effort was the highest one. At the same time, everyone bidding arbitrarily close to 0 is not sustainable in equilibrium, because overbidding the highest bid would grant victory, at a very small cost, thus deviations would be profitable, once again. Therefore, no pure strategy NE exists. For the analysis of NE in mixed strategies see Baye, Kovenock and De Vries (1996) in complete information with identical costs, and Amann and Leininger (1996) in incomplete information with two contestants. With sufficiently many contestants and prizes, Olszewski and Siegel (2016a) approximate the equilibrium behavior, even for asymmetric contests with incomplete information that are typically difficult or impossible to solve. One of the most prominent applications of the APA model is to lobbying; see Baye, Kovenock and De Vries (1993), who show that a designer interested in maximizing the sum of bids might benefit from excluding lobbyists valuing the prize the most from participating in the contest (the so-called exclusion principle). The family of APAs has been generalized by Siegel (2009), who only requires that, conditional on losing or winning, it is better to do so with a lower investment.

2.1.2 The difference-form CSF (Hirshleifer, 1989)

In this CSF the winning probability depends on the difference between a contestant's effort and a measure of the other contestants' efforts. In the special case of two contestants, the winning probability depends on the difference between contestants' efforts. Analytically,

$$p_i = F_i(G_i - G_j), \quad i, j = 1, 2, \quad i \neq j \quad (6.4)$$

Hirshleifer motivates this CSF, saying it captures "the tremendous advantage of being even just a little stronger than one's opponent" (Hirshleifer, 1991, p. 131). The problem of this



CSF is clear when assuming differentiability. In this case, the first-order conditions tell us that, since $p_2 = 1 - p_1$, in an interior equilibrium

$$F'_1(G_1 - G_2)V_1 - 1 = 0 \text{ and } F'_1(G_1 - G_2)V_2 - 1 = 0 \quad (6.5)$$

where F' is the derivative of F . Equation (6.5) implies that – if valuations are different – only the contestant with the highest valuation exerts positive effort in a pure strategy NE (Baik, 1998). Che and Gale (2000) proposed a non-differentiable special case of (6.4), namely

$$p_1 = \max \left\{ \min \left\{ \frac{1}{2} + s(G_1 - G_2), 1 \right\}, 0 \right\} \text{ with } p_2 = 1 - p_1 \quad (6.6)$$

in which the equilibrium does not necessarily occur in the region where the CSF is differentiable. Note that when $s = 0$ the CSF does not change with effort, and when $s \rightarrow \infty$ any arbitrarily small bettering of rivals' effort completely changes the outcome of the contest, and in fact (6.6) boils down to the all-pay auction (6.3). Despite this CSF has the problem spotted above – namely, at most one contestant exerts positive effort in pure strategy NE – it allows computation of mixed strategy NE.⁹

All these CSFs belonging to the family of (6.4) have the additional problem that the probabilities of winning depend on the unit of measurement of efforts (dollars or euros, minutes or hours, hundreds or thousands, etc.); that is, the probability of winning is not homogeneous of degree 0 (HDZ).¹⁰ Alcalde and Dahm (2007) proposed a CSF that maintains the idea of depending on the difference of efforts and at the same time it is HDZ given that $G_j \geq G_{j+1}$,

$$p_i = \sum_{j=i}^n \frac{G_j^\alpha - G_{j+1}^\alpha}{j \cdot G_1^\alpha}, \forall i \in N \text{ with } G_{n+1} = 0 \quad (6.7)$$

In what follows we will see another attempt to reconcile difference-form CSF and HDZ, namely (6.10) and (6.12).

2.1.3 The ratio-form CSF (Tullock, 1980) and its extensions

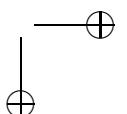
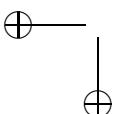
In this case the winning probability for a contestant equals the contestant's effort over the sum of all contestants' efforts. Namely,

$$p_i = \frac{G_i}{G_1 + G_2 + \dots + G_n} \quad (6.8)$$

This CSF is also known as lottery-CSF, since it equals the probability of winning the lottery with $\sum_{j=1}^n G_j$ identical tickets when you hold G_i tickets. A good property of this CSF is that it is HDZ, thus the winning probabilities are not sensitive to changes in the unit of measurement of efforts. Unfortunately, this CSF is discontinuous or undefined in $\mathbf{0}$. This, however, is a common property of HDZ-CSF (see Corchón, 2000). The usual assumption is that when all efforts equal 0, the probability of winning equals a constant $k \in [0, 1]$, and since this situation

⁹ If the cost function was strictly convex, e.g., G_i^α with $\alpha < 1$, then a pure strategy NE with positive effort exerted by both contestants exists.

¹⁰ A function $x = f(\mathbf{y})$ where $x \in \mathbb{R}$ and $\mathbf{y} \in \mathbb{R}^m$ is HDZ if $f(\mathbf{y}) = f(\lambda \mathbf{y}) \forall \lambda \neq 0$.



is never reached in equilibrium of the simultaneous complete information contest, it does not affect the outcome of the game.¹¹ Modeling a contest with the lottery-CSF is highly common and tractable. Therefore, in Section 7 we focus on this model when computing the equilibrium and reviewing results on optimal contest design.

Dixit (1987) proposed a natural generalization of (6.8), also known as logit-CSF:

$$p_i = \frac{\phi(G_i)}{\sum_{j=1}^n \phi(G_j)} \quad (6.9)$$

An interpretation of $\phi(G_i)$ is that it measures the impact of G_i in affecting the outcome of the contest. Thus, the ratio (6.9) measures the relative impacts of i 's effort on aggregate impacts of all contestants' efforts. Several CSFs proposed in the literature are special cases of (6.9). For instance, $\phi(G_i) = G_i^\epsilon$, which we will refer to as the Tullock-CSF, was proposed by Gordon Tullock in 1980. In this case, ϵ determines the returns to effort, and it is often interpreted as a noise parameter, which shows how much of a greater effort than your rival's is transformed into a greater probability of winning. In other words, the noise can be interpreted as any stochastic factor not in the hands of the contestants, such as luck, or limited observability of efforts by the contest organizer. If $\epsilon \rightarrow \infty$, the noise is absent, and an arbitrarily small overtaking of your rival's effort will make you win with certainty (the APA case). If $\epsilon = 0$, the noise is maximum; that is, the contest is a stochastic process giving to everyone $1/n$ of probability of winning, regardless of effort. If $\epsilon = 1$, the noise is intermediate, and (6.8) results.

Amegashie (2006) proposes a different way to model the noise in the contest: $\phi(G_i) = G_i + k$. The greater the noise k , the less the outcome of the contest depends on efforts.

Hirschleifer (1989) proposes a CSF that complies both with (6.9) and (6.4): $\phi(G_i) = e^{kG_i}$, with $k > 0$. In fact, it can be written as difference-form CSF in the following way

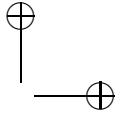
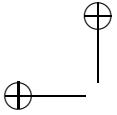
$$p_i = \frac{1}{\sum_{j=1}^n e^{k(G_j - G_i)}} \quad (6.10)$$

This CSF lacks concavity in G_i , and the best reply functions are straight lines of a 45° slope up to a point of discontinuity. As shown by Hirschleifer (1989), equilibria are either in mixed strategies or there are corner equilibria (if contestants' valuations are sufficiently asymmetric). These facts undermine its tractability. In Corchón (2007) the reader can find an overview of existence and uniqueness of NE as well as comparative statics when the CSF is in difference-form and valuations are identical, and when valuations are possibly asymmetric and the CSF is the lottery-CSF. Local properties of comparative statics for the logit-CSF with asymmetric valuations are studied in Acemoglu and Jensen (2013).

The lottery-CSF admits extensions where efforts affect probabilities of winning asymmetrically, e.g.,

$$p_i = \frac{\alpha_i G_i}{\alpha_1 G_1 + \alpha_2 G_2 + \dots + \alpha_n G_n} \quad (6.11)$$

¹¹ In fact, if every contestant plays $G_i = 0$, any arbitrarily small deviation to $G_i = \varepsilon > 0$ would grant victory at a negligible cost.



where α_i is the weight of the effort of player i . The different weights can be interpreted as different impacts of efforts, or as an unevenly leveled playing field due, for instance, to a bias, an handicap, or a judging committee biased towards some contestants. Brown (2011) make use of (6.11) to make the case that the presence of an outstanding contestant (a “superstar”) is associated with lower performance. Dahm and Porteiro (2008) also make use of (6.11) to model more accurate information of a political decision-maker lobbied by competing interests and investigate whether bias in the direction of the correct decision improves political decisions.

An CSF trying to unify the good properties of the lottery and of the difference forms is proposed by Beviá and Corchón (2015). When $n = 2$, their CSF equals

$$p_i = \alpha + \beta \frac{G_i - sG_j}{\sum_{j=1}^2 G_j} \quad (6.12)$$

where α is the part of probability inelastic to efforts – as the noise ϵ in (6.9) with $\phi(G_i) = G_i^\epsilon$ discussed above – β measures the impact of the relative efforts, and s is how the rival’s effort negatively affect i ’s probability of winning. To guarantee that $p_1 + p_2 = 1$ condition $2\alpha + \beta(1 - s) = 1$ is imposed, and to guarantee non-negative probabilities the max-min operator as in (6.6) is assumed. Adding and subtracting sG_i to the numerator of (6.12), we obtain

$$p_i = \alpha - \beta s + \beta(1 + s) \frac{G_i}{\sum_{j=1}^2 G_j} \quad (6.13)$$

therefore this CSF is an affine transformation of the ratio CSF (and they coincide when $\alpha = 0$ and $\beta = 1$, so that $s = 0$ by condition $2\alpha + \beta(1 - s) = 1$). Beviá and Corchón provide a necessary and sufficient condition for existence of an NE when $n = 2$, which is $\alpha + \beta \leq 1.5$ when valuations are identical. Moreover, the NE is unique. When $n > 2$ and valuations are identical, a sufficient condition for existence of a NE is $n \geq n(\alpha + \beta) - 1$. They also find sufficient conditions for the existence of an NE when agents have non-identical valuations.

Note that the tools to analyze games of strategic complementarities or substitutabilities do not apply to contests.¹² For instance, in the case of (6.11) with $n = 2$ and normalizing $\alpha_1 = 1$ the best reply function of contestant 1 is

$$G_1 = \sqrt{V\alpha_2 G_2} - \alpha_2 G_2, \quad (6.14)$$

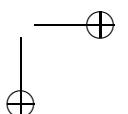
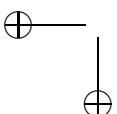
which is first increasing (strategic complements) and then decreasing (strategic substitutes). It is easy to prove that this property carries over to logit-CSF, under some very mild conditions (see Dixit, 1987).

2.2 Other Contest Success Functions

2.2.1 Ties

The CSFs seen so far do not admit intermediate outcomes; either the contest is won or lost. Yet, in several of the applications highlighted in the Introduction, ties are natural and likely

¹² The actions of players are strategic substitutes (complements) when they mutually offset (reinforce) one another. See Bulow, Geanakoplos and Klemperer (1985).



outcomes of the contest. There are wars that reach an impasse without a clear winner, like the Korean War (1950–53) or the Iran and Iraq (1980–88) Wars. Trials in Anglo-Saxon countries might end up with a null verdict or a “hung jury.” Ties also play an important role in sports such as football, chess and cricket. Finally, in some procurements the prize (or work) might not be allocated when none of the contestants meet the minimal quality requirements.

Despite their relevance in contests, the literature started considering ties only recently. Here we consider $n = 2$ to avoid having to consider ties of all the possible subgroups of n contestants. The first contribution is Blavatskyy (2010), which shows that under a set of axioms inspired by the work of Skaperdas (1996) – which will be discussed in the next section – the probability of ties is

$$\frac{1}{1 + \alpha_1 G_1^r + \alpha_2 G_2^r} \quad (6.15)$$

where α_1, α_2 and r are real and positive numbers. The problem of such a CSF is that, as noted by Peeters and Szymanski (2012), it is arguably implausible that the probability of ties goes to zero as efforts of both contestants are identical and large. They present a CSF where the probability of ties is given by the relative difference, with the goal of empirically testing it. Jia (2012) axiomatizes a CSF where the probability of ties is

$$\frac{G_1 G_2 (c^2 - 1)}{(G_1 + cG_2)(G_2 + cG_1)} \quad (6.16)$$

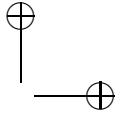
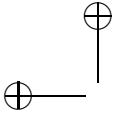
where $c > 1$. Jia shows that there is a unique NE in pure strategies when $c < 3$. Finally, Yildizparlak (2013) presents a CSF that admits ties, and applies it to four European football leagues (German, Spanish, French and Italian) with positive results. In his work, the probability of ties is given by

$$1 - \frac{\sum_{j=1}^2 \phi(G_j)^k}{(\sum_{j=1}^2 \phi(G_j))^k} \quad (6.17)$$

where $\phi(\cdot)$ is increasing, concave and differentiable, and $k > 1$ is the parameter interpreted as the propensity to ties. When $k = 1$ the CSF boils down to the logit-CSF. Yildizparlak shows that there exists a unique NE in pure strategies when $k < 3$. Curiously, the value of k , which in theory maximizes efforts, $k = 1.44$, is very close to the one estimated in the four European leagues, which is close to 1.5.

The possibility of ties in APA has been analyzed by Gelder, Kovenock and Roberson (2015). They characterize the Nash equilibria of a symmetric complete information setting with two contestants where a tie occurs if neither player outbids the other by strictly more than a fixed amount, $\delta \geq 0$. When players tie, they receive an identical fraction of the prize, $\beta \in [0, 1]$. Thus, the CSF can be written as follows

$$p_i(G_i, G_j) = \begin{cases} 1 & \text{if } G_i - G_j > \delta \\ \beta & \text{if } |G_i - G_j| \leq \delta \\ 0 & \text{if } G_i - G_j < -\delta \end{cases} \quad (6.18)$$



2.2.2 Mechanism design

Polishchuk and Tonis (2013) present a novel view to contest design under informational asymmetry among contestants. The designer picks the CSF. By means of the revelation principle, they retrieve the optimal CSF. For a given distribution of types, the CSF that maximizes expected aggregate efforts is the logit-CSF, whereas for other distribution of types they find that the optimal CSF is the additive logarithmic

$$p_i = \frac{1}{2}(\log G_i - \log G_j) + \frac{1}{2} \quad (6.19)$$

the difference-form CSF of Che and Gale (2000), or the following CSF that combines the additive and the difference CSF, and that is similar to the one proposed by Beviá and Córchón (2015):

$$p_i = \frac{2\phi(G_i) - \phi(G_j)}{\sum_{j=1}^2 \phi(G_j)} \quad (6.20)$$

When requiring ex post optimality, the optimal CSF has similarities with the famous Vickrey-Clarke-Groves mechanism. This is not surprising, given that contests can be regarded as the allocation of a public good.

A problem with these results is that besides the equilibrium where agents truthfully report their valuations, others might coexist where agents lie. Furthermore, the optimal CSF depends on the distribution of types, which is usually unknown. Still, much has to be done to fully understand limits and advantages of the mechanism design viewpoint to discover new CSFs.

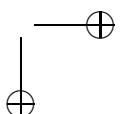
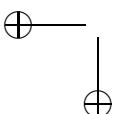
3 MICROFOUNDATIONS OF CSF

So far we just assumed the CSF's functional form, which could be more or less plausible. The literature provided several ways to microfound some standard CSFs. Microfoundations are also useful to discover new CSFs.

3.1 Stochastic Performance

In this viewpoint, started by Dixit (1987) and Hillman and Riley (1989), it is assumed that performance is made of two factors: the effort of the contestant, and a stochastic component that we will denote by ϵ_i . For simplicity, we assume $n = 2$. Thus, contestant i wins the contest if and only if $\epsilon_i G_i > \epsilon_j G_j$. Note that the all-pay auction is a special case (i.e., if $\epsilon_i = \epsilon_j$).

The general result is given by Jia (2008), which proves that if the stochastic components follow an inverse exponential distribution the resulting CSF is the logit-CSF. The works by Fullerton and McAfee (1999) and Baye and Hoppe (2003) also offer microfoundations for the logit-CSF in the cases of innovations and patents. These foundations, although natural, sharply rely on the particular shape of the cumulative distribution function of the stochastic component, of which we know very little.



3.2 Axiomatic

The axiomatic approach focuses on properties that characterize the CSF. Here, the seminal work is Skaperdas (1996).¹³ Skaperdas shows that the logit-CSF is obtained under the assumptions of (1) imperfect discrimination (i.e., the probability of winning is always positive if effort is positive); (2) monotonicity (i.e., the probability of winning increases in effort); (3) anonymity (the probability of winning depends on an agent's effort, not on an agent's identity); and (4) a form of irrelevance of independent alternatives. If the extra property of HDZ is added, the Tullock-CSF is obtained. In a subsequent work by Clark and Riis (1998a), a neat proof generalizes Skaperdas' axiomatization to asymmetric CSF. A paper by Vesperoni (2013) treats the case of multiple prizes, and axiomatizes an appropriate CSF for this context. Münster (2009) extends Skaperdas' axiomatization of the logit-CSF to group contests. Cubel and Sanchez-Pagés (2014) axiomatize the difference-form CSF for the case in which the contestants are groups.

3.3 Contest Designer

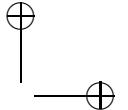
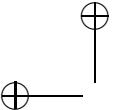
As mentioned above, some contests, like sports and public procurements, are designed and managed by one or several people. We refer to them as *the designer*. In this context, the CSF is, to some extent, the direct consequence of the actions of the designer. An early work in this field is Epstein and Nitzan (2006). They assume $n = 2$ and that the objective of the designer is a weighted average of the social welfare and of contestants' efforts. This latter has a two-fold interpretation: first, the efforts could improve the quality of the prize, as the plans for the Olympic Games improve the quality of the Olympic Games themselves, and second, efforts could be interpreted as monetary transfers from contestants to the designer. Epstein and Nitzan study the conditions under which the designer prefers to organize a contest with a Tullock-CSF or an all-pay auction.

In Corchón and Dahm (2010) the designer could be of various types, unknown to contestants. In turn, contestants might ignore the bias in the designer's preferences (towards one contestant, or another), her way of working, her ideology, etc. The CSF is simply the best reply of the designer as perceived by contestants. For instance, consider $n = 2$ and the designer's utility is

$$U_1 = (1 - \theta)\phi(G_1) \quad \text{and} \quad U_2 = \theta\phi(G_2) \quad (6.21)$$

where U_i is the designer's utility if contestant i wins, ϕ measures the impact on the designer's utility, and θ is uniformly distributed in $[0, 1]$. For given efforts, the probability that contestant 1 wins is the probability that $U_1 > U_2$, which is $\phi(G_1)/(\phi(G_1) + \phi(G_2))$, i.e., the logit-CSF. In this case, we say that the logit-CSF could be *rationalized*. Corchón and Dahm show that when $n = 2$ every CSF could be rationalized, whereas when $n > 2$ none of the known CSFs could be rationalized. This is because in the standard CSFs the probability of winning depends symmetrically on the aggregate efforts of the others. Yet, the competition is very much local, such as in some industrial organization models (e.g., the Salop model) where a player's payment depends on the payments of the neighboring players.

¹³ The interested reader can find explanation and discussion of this microfoundation in Corchón (2007).



In a subsequent work Corchón and Dahm (2011) consider a designer who cannot commit to a CSF, and can only allocate probabilities of winning once efforts are exerted, which by itself is a CSF. They suppose that the designer maximizes a constant elasticity of substitution objective function, where the arguments are the efforts,

$$W(\mathbf{p}, \mathbf{G}) = \begin{cases} \left(\sum_{i=1}^n (p_i V_i - G_i)^{1-r} \right)^{1/(1-r)} & \text{if } r \neq 1 \\ \sum_{i=1}^n \ln(p_i V_i - G_i) & \text{if } r = 1 \end{cases} \quad (6.22)$$

If $r = 1$ the first-order conditions give

$$(p_i V_i - G_i) V_j = (p_j V_j - G_j) V_i, \quad \forall i \in N \quad (6.23)$$

Solving this system we obtain

$$p_i = \frac{1 - \sum_{j=1}^n (G_j/V_j)}{n} + G_i/V_i \quad \forall i \in N, \quad (6.24)$$

which is a CSF linear in the differences, such as (6.2). The conclusion carries over to any other r . Finally, they show that the logit-CSF cannot be rationalized by any designer who maximizes a function of contestants' efforts. Yet, the logit-CSF is obtainable when the designer has an objective function à la Kahneman-Tversky (1979).

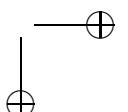
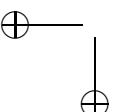
3.4 Other Foundations

Corchón and Dahm (2010) assume that contestants bargain over the outcome of the contest as in lawsuits for the ownership of resources. In this case it is best to interpret p_i as the share of the resources allocated to contestant i . Inspired by a result in Dagan and Volij (1993), they microfound the logit-CSF as the result of the (asymmetric) Nash bargaining solution where efforts are the weights of each agent. Also, they establish a connection to the bargaining problems with claims where efforts induce “aspiration” $\phi_i(G_i)$. They find that the proportional problem to the bargaining problem induces the logit-CSF while the claim-egalitarian solution induces the difference-form CSF and relative claim-egalitarian solution induces the CSF (6.7).

Skaperdas and Vaidya (2012) microfound the ratio and the difference-form CSF considering a setting where contestants produce “evidence” from which a Bayesian judge infers the guilt of a defendant.

4 DYNAMIC CONTESTS

The contest models analyzed so far do not take into consideration the dynamic component of several real-life contests. For instance, many television contests have an eliminatory stage that leads to the final among a reduced number of participants. The same occurs in several football and tennis leagues where participants play in eliminatory rounds. These are called *eliminatory*



contests. The seminal paper on elimination contests is Rosen (1986).¹⁴ It could also occur that the contest is made up of several stages, contestants obtain an outcome in each stage, and the contest is won by the contestant with the greatest overall outcome, or with the greatest stage outcome. The former occurs in leagues (football, basketball) or in the election of a candidate for the Republican or Democratic parties. Wars often consist of several consecutive battles. The latter occurs in innovation races, where firms produce innovations of stochastic value in each period, and the firm that manages to produce the innovation of the greatest value wins the contest; see Taylor (1995). These contests are usually called *races*.

A survey of the literature is found in Konrad (2012). It seems reasonable to expect that, in a race with a unique final prize where one contestant accumulates a significant advantage, all contestants lessen their efforts; the advantaged contestant because she needs little effort in order to win, and the contestants lagging behind because they would need a great effort to catch up. This is what Konrad calls the “discouragement effect.” A consequence of this is called the “New Hampshire effect”; the first and preliminary results of the elections of a candidate for the USA presidency are a good predictor of the winner; see Klumpp and Polborn (2006). This is due to the discouragement effect that magnifies the advantage of the winners of the early stages of the contest. This is generalized in the so-called Matthew effect, coined by the famous sociologist Robert K. Merton, which takes its name from the biblical Gospel of Matthew, “the rich get richer and the poor get poorer,” where wealth is meant metaphorically to refer to fame or status. Thus, an already famous scholar tends to receive more citations than less known scholars, despite similar quality works.

But the discouragement effect, important as it is, is not a general property of dynamic games. For instance, if the loser is concerned with the magnitude of the defeat, a first battle defeat could actually incentivize the loser to exert more effort in subsequent battles, in order to avoid a dishonorable defeat; see Sela (2011). Möller (2012) and Beviá and Corchón (2013) consider two-period contests where the first period has a positive effect on the probability of winning in the second period, and players have an incentive to exert effort. Therefore, the discouragement effect only occurs when the initial difference between the two contestants is sufficiently large. The Matthew effect occurs with the α s in (6.11), but not with the received share of the prize. Hence, contestants who are initially advantaged by the CSF, tend to be even more advantaged in the second period, and in turn this incentivizes them to exert less effort in the second period, which reduces their advantage. A generalization of these issues to models with more than two periods has been pursued by Luo and Chie (2016).

Besides scheduling the stages of the contest sequentially, several real-life contests are simultaneous subcontests in nature, where the winner is the one winning the vast majority of subcontests. Fu, Lu and Pan (2015) establish that in multi-battle team contests where the contest success function is homogeneous of degree zero (which nests both the logit-CSF and the APA), the outcomes of past battles do not affect the outcome of future battles, and that aggregate effort and contest outcome are not affected by the temporal structure of the contest – simultaneous or (partially) sequential battles – or its feedback policy. Besides playing simultaneously or sequentially across subcontests (battlefields), one can think of contestants playing sequentially within the same contest. This case is analyzed by Morgan (2003), who finds that sequential contests are ex ante Pareto superior to simultaneous contests, and that if contestants can choose between simultaneous or sequential contests before type

¹⁴ On elimination contests see also the subsequent works, Amegashie (1999b) and Konrad and Gradstein (1999).

realizations, the latter is chosen in any subgame perfect equilibrium. Leininger (1993) covers the case where contestants choose the timing of the contest after type realizations, and the main finding is that contestants agree to move in a particular sequential order. Extension of these results to the asymmetric information setting has been carried out by Fu (2006). The incentives to precommit to effort are analyzed in Dixit (1987): in a two-player contest, he finds that the stronger (weaker) contestant will want to precommit to a higher (lower) level of effort than the Nash level without precommitment. Baik and Shogren (1992) extend Dixit's model to allow for endogenous order of moves, and find that the weak contestant plays first.

5 ASYMMETRIC INFORMATION

Sometimes the assumption that contestants know each other well is reasonable. In the final of the 2015 UEFA Champions League, it was widely well known that Barcelona and Juventus had very different strengths (1€ on Juventus' victory yielded 5.88€, 1€ on Barcelona's victory yielded 1.63€).¹⁵ At other times, the assumption of complete information is questionable; in job interviews candidates often ignore who the other candidates are, in warfare the amount of resources or the military technologies owned by each party are often private information, or in a research contest scientists have often a very limited knowledge of the set of contestants. Predictions of models sharply change when dropping the complete information assumption.

Einy et al. (2010) provided an example where the Bayesian equilibrium does not exist, despite the payoff functions given that types are quasiconcave.¹⁶ Einy et al. (2015) show that in the special case of logit-CSF, a Bayesian equilibrium in pure strategies exists. When a contestant has greater information than her rival the equilibrium is unique, and the contestant with greater information expects greater payoff, although wins the prize with lower probability. Thus, similar to the case of leveling the playing field, the contestant with a competitive advantage makes the most out of it, and exerts less effort.

Corchón and Yıldızparlak (2013) study war as a game where the declaration of the war itself signals information about the type of contestant. Despite the fact that contestants could transfer resources to level out the inequalities and so incentivize peace, war is sometimes unavoidable and could occur with very little asymmetric information.¹⁷ See Jackson and Morelli (2011) and Baliga and Sjostrom (2011) for surveys on the theory of conflicts applied to warfare.

A recent literature endogenizes the information. Suppose that the designer knows contestants' types, but each contestant *a priori* knows only her own type, and not her rival's. This is the case in grant contests, job interviews, chess championships, and many others. If the designer maximizes expected aggregate efforts, should she disclose contestants' information on types to the others? While the answer for a revenue-maximizing auctioneer is always positive (see the "linkage principle" by Milgrom and Weber, 1982), in contests this is not always the case. In fact, Serena (2016) shows that when the distribution of types is binary

¹⁵ Source: <http://www.oddsportal.com/soccer/europe/champions-league-2014-2015/results/>.

¹⁶ This is due to the fact that the sum of quasiconcave functions need not to be quasiconcave, thus the expected utility is not quasiconcave, and best reply functions are not convex valued.

¹⁷ Under complete information, the transfers yield peace unless the initial distribution of resources is greatly heterogeneous; see Beviá and Corchón (2010).

(V_h with probability p , $V_l \leq V_h$ with probability $1 - p$), the designer is strictly better off precommitting to fully disclose when $p \in (0.5, 1)$, and to fully conceal when $p \in (0, 0.5)$. The intuition relies on the following simplified reasoning. When p is very high, concealment would be severely detrimental in that a low type would not only believe she was up against a high type, but also believe that her high-type rival believes he has to fight hard because he is likely to be against another high type. This yields a disproportionate discouragement of the low type. On the other hand, when p is very low, concealment would be very beneficial in that a high type would believe he is up against a low type who believes she is up against another low type and thus does not give up exerting effort, which has a positive impact on the high type's effort. Note that this second effect is possible only under asymmetric information.

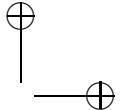
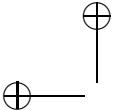
The role of information in a common and uncertain value contest is considered in Wärneryd (2003). He finds that if the two contestants are evenly informed about the value of the prize they are competing for, the equilibrium efforts can be greater than if only one of the two contestants is fully informed about the value of the prize.

6 CONTEST AMONG GROUPS

There are plenty of real-life situations where contestants are groups, rather than individuals. Groups of scientists join forces to make a breakthrough and win a research contest. Members of political parties work together in order to win the elections. The probability of a group to win the contest depends on the aggregate effort of the members of the group. This creates a free-rider problem. Mancur Olson (1985) conjectured that if contributions of group members are voluntary, small groups are more capable of coping with the free-rider effect. This is referred to as the group size paradox: smaller groups are more effective in group contests. In general, this paradox is not true – see Corchón (2007) and the original references therein – because when a group admits a new member, two effects occur: on the one hand, the effort of the old members decreases; on the other hand the effort of the new member is added. The second effect might dominate, as in Cournot models, where entry reduces incumbents' output, but increases total output. Nitzan and Ueda (2009) show conditions for the existence (non-existence) of the group size paradox.

Katz, Nitzan and Tversky (1990) analyze group contests where groups vary in the number of members and the prize is a pure public good. They find that when all members are identical, all groups exert the same aggregate effort regardless of asymmetries in group size, and that groups with greater valuations V_i tend to exert more effort, and thus win with greater probability. Their analysis is run under (6.8), where G_i is the sum of the efforts of group i 's members. Nti (1998) generalizes the CSF of Katz et al. (1990) to (6.9).

A strand of the literature analyzes what happens when groups can design internal rules to incentivize efforts – such as internal reward schemes; see Nitzan and Ueda (2011). In Lee and Kang (1998), intra-group sharing rules are determined, and then individual outputs are chosen – under (6.8) – and the contest is associated with externalities in that each member's cost of rent-seeking is negatively or positively affected by the aggregate effort. They find that if this effect is positive, the rent is more dissipated in the collective contest than in the individual contest. A further step ahead has been made thanks to the contribution of Vázquez-Sedano (2014), who applies mechanism design to the group contest sharing rule.



138 *Handbook of game theory and industrial organization: applications*

A key ingredient of group contests is how efforts of the group members are aggregated. Three cases are the most prominent in the literature:

- best-shot contests, where group performance depends on the best performer within the group; see, for instance, Chowdhury et al. (2013) for the Tullock contest, and Barbieri, Malueg and Topolyan (2014) for the APA;
- contests where contestants' efforts are perfect substitutes; see, for instance, Katz et al. (1990) and Baik (2008) for the Tullock-CSF, and Baik, Kim and Na (2001) and Topolyan (2014) for the APA;
- weakest link contests, where group performance depends on the worst performer within the group; see, for instance, Lee (2012) for the Tullock contest, and Chowdhury, Lee and Topolyan (2016) for the APA.

Combinations of the above three ways of aggregating contestants' efforts can be found in Kolmar and Rommeswinkel (2013) and Chowdhury and Topolyan (2015, 2016).

7 EQUILIBRIUM AND OPTIMAL CONTEST DESIGN

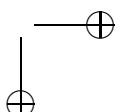
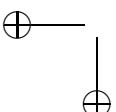
Several CSFs have been discussed in this survey. In this section we focus on the lottery-CSF, one of the most prominent and tractable ways of modeling contests, and its generalizations. We start with the analysis of the equilibrium of such a game, and we then draw conclusions on how to optimally design such contests. For the optimal design of contests, the broadly prevailing fashion is to assume that the contest designer maximizes the sum of contestants' efforts. That is, the designer aims to stimulate competition among contestants, and equally benefits from every contestant's effort. Serena (2017) proposes as an alternative the maximization of effort of the winner: the designer of an architectural contest only benefits from the effort exerted behind the winning entry (which is the only project that will eventually be implemented). Despite other alternative objective functions present in the literature,¹⁸ we mostly focus here on the maximization of the sum of efforts and of the effort of the winner.

By concavity of the lottery-CSF in own effort, the second-order conditions of utility maximization are satisfied, and thus the NE can be computed from first-order conditions. Order contestants by prize valuation $V_1 \geq V_2 \geq \dots \geq V_n > 0$. Only the first $m \leq n$ contestants exert positive efforts in the unique equilibrium, and for them,

$$p_i^* = 1 - \frac{m-1}{V_i \left(\sum_{j=1}^m \frac{1}{V_j} \right)}, \quad G_i^* = \frac{m-1}{\sum_{j=1}^m \frac{1}{V_j}} \left[1 - \frac{m-1}{V_i \left(\sum_{j=1}^m \frac{1}{V_j} \right)} \right], \quad \text{and } \Pi_i^* = \left[1 - \frac{m-1}{V_i \left(\sum_{j=1}^m \frac{1}{V_j} \right)} \right]^2 \quad (6.25)$$

m is characterized by the greatest i such that $G_i^* \geq 0$ (see Fullerton and McAfee, 1999). As for contestants $i = m+1, \dots, n$, in equilibrium $G_i^* = p_i^* = \Pi_i^* = 0$. In other words, they

¹⁸ Falconieri, Palomino and Sákovics (2004), Palomino and Sákovics (2004), and Vrooman (2012) consider the maximization of competitive balance, understood as the uncertainty of the contest outcome: uncertainty excites the interest of the audience of sport events. Azmat and Möller (2009) consider a contest designer who wants to attract participation – as in online communities where users contribute to the content.



quit, because competition is too tough and their prize valuation not sufficiently high. Several conclusions can be drawn from (6.25). First, note that the sum of efforts simplifies to

$$\sum_{i=1}^n G_i^* = \frac{m-1}{\sum_{j=1}^m \frac{1}{V_j}}$$

and thus it “closely approximates the harmonic mean of individuals’ valuations as the number of contestants increases” (see Hillman and Riley, 1989). Under a mild technical condition on the distribution of V_i s,¹⁹ Fullerton and McAfee (1999) show that, if the designer can set an entry fee E , the total cost of achieving a desired sum of efforts (including the collected entry fees) is minimized at $m^* = 2$. If the designer cannot set and collect entry fees, Fang (2002) shows that exclusion of contestants is not beneficial to a designer interested in maximizing the sum of efforts.²⁰ This result is due to the fact that excluding contestants has two effects: (1) the effort of the excluded contestant is not exerted any longer, and (2) individual efforts change (increase or decrease according to the type of excluded contestant). Fang’s results show that (1) is always stronger than (2) in affecting the sum of efforts. On the other hand, consider a designer who maximizes expected winning effort; that is, a designer interested in maximizing the quality of the winning entry. Then (1) does not affect per se the expected winning effort, whereas (2) might increase the expected winning effort (necessary and sufficient conditions for this to happen are provided by Serena, 2017).

When contestants have identical valuations $V_i = V \forall i \in N$, it is easy to see that $G_i^* > 0 \forall i \in N$ (i.e., no one quits: $m = n$), and in the unique NE:

$$p_i^* = \frac{1}{n}, \quad G_i^* = \frac{n-1}{n^2}V, \quad \text{and } \Pi_i^* = \frac{V}{n^2} \quad (6.26)$$

From (6.26) one can see that individual effort increases with the prize and decreases with the number of contestants. Thus, a designer who maximizes the effort of the winner (and cannot set and collect entry fees) should exclude contestants so as to have a two-player contest. On the other hand, a designer who maximizes the sum of efforts (i.e., $\frac{n-1}{n}V$), benefits from increasing n , and thus should be concerned with ways to stimulate participation and advertize the contest. Furthermore, $\lim_{n \rightarrow \infty} \sum_{j \in N} G_j^* = \lim_{n \rightarrow \infty} V$; thus, if both V and G_i ’s are money, contestants tend to spend all in all an amount of money that equals the prize at stake, as the number of contestants grow large. Tullock named this property “full rent dissipation.”

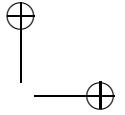
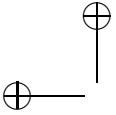
We conclude this section by discussing the role of noise in the CSF and of leveling the playing field.

7.1 Noise

As discussed, a tractable model to include a variable amount of noise in the contest is (6.9) with $\phi(G_i) = G_i^\epsilon$. Noise in the winner selection process is naturally embedded in many real-life applications. The designer of a research contest has limited time or money to carefully

¹⁹ The technical condition is that $\frac{m}{V_m} \left(\sum_{j=1}^m \frac{1}{V_j} \right)$ is non-decreasing, which is, for instance, satisfied if V_i is constant, if there are constant increments to valuations, or proportional increments to valuations.

²⁰ This result is in sharp contrast with the above-mentioned exclusion principle; see Baye et al. (1993).



evaluate every little detail of the submitted projects, thus there is a probability of making the mistake of not selecting the highest effort as winner of the contest. In sports competitions, even the weakest player usually has a positive chance of winning thanks to the inherent stochasticity of sports competitions, for instance if a player gets sick on the day of the match, or if an inexperienced archer happens to be lucky at an archery contest. The noise is either exogenous (good or bad luck in sports competitions) or – to some extent – endogenous (amount of time spent on evaluating projects by the designer of a research contest). When it is endogenous, it is relevant to understand whether more or less noise is beneficial to the contest designer. For simplicity, a common assumption is that $\epsilon \in [0, 1]$, which is always sufficient to make CSF concave and thus the first-order approach valid.²¹ In particular, the unique NE under $\epsilon \in [0, 1]$ of the symmetric Tullock contest is

$$p_i^* = \frac{1}{n}, \quad G_i^* = \frac{\epsilon(n-1)V}{n^2}, \quad \text{and } \Pi_i^* = \frac{V(n-\epsilon(n-1))}{n^2} \quad (6.27)$$

Note that (6.27) generalizes (6.26). When the number of contestants grows large (i.e., $n \rightarrow \infty$), the property of “full rent dissipation” fails to hold if the contest is sufficiently noisy (i.e., $\epsilon < 1$), since total effort tends to ϵV . The negative effect of the noise carries over to a contest where the designer maximizes the effort of the winner. Additionally, contestants are worse off by the presence of noise, since individual and total utilities decrease in ϵ . Thus, the noise is detrimental in a contest among identical contestants. Yet, in a contest among heterogeneous contestants, the noise could have the positive effect of giving hope to the low type, which in turn stimulates efforts of the high type and is beneficial to the contest organizer. An easy way to see this positive effect of the noise is to consider a two-player contest with heterogeneous valuations: $V_1 > V_2 > 0$. The unique equilibrium is

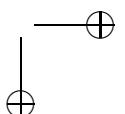
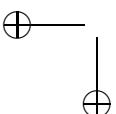
$$G_i^* = \epsilon \frac{V_1^\epsilon V_2^\epsilon}{(V_1^\epsilon + V_2^\epsilon)^2} V_i, \quad (6.28)$$

which can be proved to be concave in ϵ . Additionally, routine algebra shows that

$$\frac{\partial G_1^*}{\partial \epsilon} \gtrless 0 \iff \frac{\partial G_2^*}{\partial \epsilon} \gtrless 0 \iff \tilde{V}^\epsilon + 1 \gtrless \epsilon (\tilde{V}^\epsilon - 1) \log \tilde{V}^\epsilon, \quad (6.29)$$

where $\tilde{V} = \frac{V_1}{V_2} > 1$. Thus, consider the extremes of the interval $\epsilon \in [0, 1]$: if $\epsilon \rightarrow 0$, $\frac{\partial G_i^*}{\partial \epsilon} > 0$, whereas if $\epsilon = 1$, $\frac{\partial G_i^*}{\partial \epsilon} < 0$ if \tilde{V} is larger than 4.68 (or equivalently, smaller than $\frac{1}{4.68}$). In other words, if there is sufficient heterogeneity of types, the optimal noise for a designer that maximizes individual effort or sum of efforts is $\epsilon^* \in [0, 1]$. The intuition is that *too much* noise makes the probability of winning inelastic to effort and thus equilibrium efforts are small, whereas *too little* noise – if contestants are sufficiently asymmetric – gives no hope of winning to the low type, who thus exerts very little effort. Considering instead the expected

²¹ Pérez-Castrillo and Verdier (1992) find the noise threshold ϵ^* such that in a Tullock contest a pure strategy NE exists if and only if $\epsilon \in [0, \epsilon^*]$. They find that $\epsilon^* \in [1, 2]$ and it depends on contestants’ valuations. The case of $\epsilon > 2$ is tricky, and only mixed strategy NE exist; see Baye, Kovenock and De Vries (1994), Alcalde and Dahm (2010) and Ewerhart (2015).



winning effort, the same conclusion carries over and the optimal $\epsilon^{**} \in [0, 1]$, but it could be proved that this new optimal level of noise is greater than the one maximizing sum of efforts.

For general properties of the optimal ϵ including both the range for which a pure strategy exists and the range for which it does not, thus any $\epsilon > 0$, see Wang (2010).

7.2 Leveling the Playing Field

As discussed, a tractable model to include the possibility of leveling the playing field is (6.11).²² If the designer could choose α_i s so as to maximize the sum of efforts, she would *perfectly level the playing field*, that is, to give an advantage to the low types so as to make contestants equally likely to win in equilibrium, and thus maximize competition for the prize; see Franke (2012a). This is because in equilibrium the strategies in the best reply of the strong player are strategic complements. Thus raising the best reply of the disadvantaged player increases the effort of both players. Such policies do positively affect efforts in real-life tournaments; see Brown (2011) and Franke (2012b) for such evidence in golf tournaments, Calsamiglia, Franke and Rey-Biel (2013) for such evidence in Sudoku tournaments, and Levitt (1994) for such evidence in campaign expenditures in US House elections. If instead the designer could choose α_i s so as to maximize the effort of the winner, then it is optimal to leave the high-type more likely to win in equilibrium (Serena, 2017). The intuition is as follows. A perfectly leveled playing field achieves the maximum of the effort of both contestants. If instead the playing field is not perfectly leveled, in particular such that the high type is more likely to win in equilibrium, then two effects – one positive and one negative – simultaneously arise: the negative effect is that both contestants exert their non-maximum effort, and the positive effect is that the probability of winning of the high type (whose effort is the greatest) increases. Serena (2017) shows that this trade-off yields the optimality of leaving some degree of advantage to the high type in a contest where the designer maximizes the effort of the winner.

Despite that in general leveling the playing field is good for a contest designer, Brown and Chowdhury (2014) show theoretically that when sabotage is possible, leveling the playing field may increase sabotage, which is detrimental to the contest designer. They support the theory with horse racing data. They show that a handicap in horse racing (in which favorite horses carry extra weights in the saddle) works well in terms of increasing competition. However, that also increases the unwanted interruptive behavior of the jockeys such as bumping into other horses and running dangerously.

8 FURTHER TOPICS

This survey would be incomplete without mentioning some other topics analyzed by the literature. In particular, recent surveys do a commendable job of covering these topics:

1. Testing the theoretical results is an important milestone of the analysis of contests. Yet, empirical analysis of contests is somehow problematic in that efforts are not directly

²² Note that there are other ways to level the playing field besides by biasing the CSF choosing α_i s., for instance, by means of an extra prize (see Dahm and Esteve, 2016).

- observable in the field.²³ A discussion of the problems is found in Jia, Skaperdas and Vaidya (2013). For this reason, experimental results have been booming in the recent years. One of the predominant findings of experimental results on contests is a significant overbidding as opposed to the Nash equilibrium. Several explanations have been provided, mostly based on modified utility function (including, for instance, non-monetary utility from winning the contest, or preferences over payoffs relative to other contestants) or on subjects' irrational behaviors (subjects' proneness to mistakes, or subjects' judgmental bias such as the hot hand fallacy: the fallacious belief that someone who has experienced success in a random event has a greater chance of success in additional attempts). Overbidding in group contests with intra-group punishment opportunities is perhaps the sharpest one relative to theoretical predictions, as documented by Abbink et al. (2010). For a comprehensive recent survey on experimental results on contests see Dechenaux, Kovenock and Sheremeta (2015).
2. In many real-life contests, besides exerting costly effort to increase their probability of winning, players might exert costly effort to "sabotage" the rival's likelihood of winning. A recent survey on sabotage in contests is Chowdhury and Gürtler (2015).
 3. Multi-winner contests in which more than one player can win at most one prize is one area that is under-researched and emerging. Seminal contributions on multi-winner contests are Clark and Riis (1998b) and Moldovanu and Sela (2001). The latter studies the optimal prize structure in multi-winner contests, and its main findings are confirmed and generalized in large contests (i.e., with sufficiently many contestants) by Olszewski and Siegel (2016b). When the contest is not an all-pay auction, there is more than one way of formulating the probability of ending up n th in the contest, and thus win the n th prize. CSF in that area are proposed by Berry (1993) who suggests picking k winners simultaneously among $n (> k)$ players. Clark and Riis (1996) provide a sequential mechanism in which k winners are picked sequentially one by one. Chowdhury and Kim (2016) on the other hand propose a sequential mechanism in which $(n - k)$ losers are taken out sequentially one by one. See Sisak (2009) for a survey.

REFERENCES

- Abbink, K., J. Brandts, B. Herrmann and H. Orzen (2010). "Intergroup Conflict and Intra-group Punishment in an Experimental Contest Game: Aggregate Comparative Statics," *The American Economic Review*, 100(1), 420–447.
- Acemoglu, D. and M.K. Jensen (2013). "Aggregate Comparative Statics," *Games and Economic Behavior*, 81, 27–49.
- Alcalde, J. and M. Dahm (2007). "Tullock and Hirshleifer: A Meeting of the Minds," *Review of Economic Design*, 11(2), 101–124.
- Alcalde, J. and M. Dahm (2010). "Rent Seeking and Rent Dissipation: A Neutrality Result," *Journal of Public Economics*, 94(1–2), 1–7.
- Alcalde, J. and M. Dahm (2013). "Competition for Procurement Shares," *Games and Economic Behavior*, 80, 193–208.
- Amann, E. and W. Leininger (1996). "Asymmetric All-pay Auctions with Incomplete Information: The Two-player Case," *Games and Economic Behavior*, 14(1), 1–18.
- Amegashie, J.A. (1999a). "The Number of Rent-seekers and Aggregate Rent-seeking Expenditures: An Unpleasant Result," *Public Choice*, 99(1), 57–62.
- Amegashie, J.A. (1999b). "The Design of Rent-seeking Competitions: Committees, Preliminary and Final contests," *Public Choice*, 99(1/2), 63–76.

²³ An empirical test of CSFs is to be found in Hwang (2012), who proposes and tests a combination of difference and ratio CSF using data from seventeenth-century European battles and World War II.

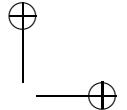
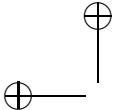
- Amegashie, J.A. (2006). "A Contest Success Function with a Tractable Noise Parameter," *Public Choice*, 126(1), 135–144.
- Anderson, S.P., J.K. Goeree and C.A. Holt (1998). "Rent Seeking with Bounded Rationality: An Analysis of the All-pay Auction," *Journal of Political Economy*, 106(4), 828–853.
- Azmat, G. and M. Möller (2009). "Competition Amongst Contests," *The RAND Journal of Economics*, 40(4), 743–768.
- Baik, K.H. (1998). "Difference-form Contest Success Functions and Effort Level in Contests," *European Journal of Political Economy*, 14(4), 685–701.
- Baik, K.H. (2008). "Contests with Group-specific Public-good Prizes," *Social Choice and Welfare*, 30(1), 103–117.
- Baik, K.H. and J.F. Shogren (1992). "Strategic Behavior in Contests: Comment," *The American Economic Review*, 82(1), 359–362.
- Baik, K.H., Kim, I.G. and Na, S. (2001). "Bidding for a Group-specific Public-good Prize," *Journal of Public Economics*, 82(3), 415–429.
- Baliga, S. and T. Sjostrom (2013). "Bargaining and War: A Review of Some Formal Models," *Korean Economic Review*, 29(2), 235–266.
- Barbieri, S., D.A. Malueg and I. Topolyan (2014). "The Best-shot All-pay (Group) Auction with Complete Information," *Economic Theory*, 57(3), 603–640.
- Baye, M.R. and H. Hoppe (2003). "The Strategic Equivalence of Rent-seeking, Innovation, and Patent-race Games," *Games and Economic Behavior*, 44(2), 217–226.
- Baye, M.R., D. Kovenock and C. de Vries (1993). "Rigging the Lobbying Process: An Application of the All-pay Auction," *The American Economic Review*, 81(1), 289–294.
- Baye, M.R., D. Kovenock and C. de Vries (1994). "The Solution to the Tullock Rent-seeking Game When R>2: Mixed-strategy Equilibria and Mean Dissipation Rates," *Public Choice*, 81(3), 363–380.
- Baye, M.R., D. Kovenock and C. de Vries (1996). "The All-pay Auction with Complete Information," *Economic Theory*, 8(2), 291–305.
- Becker, G. (1983). "A Theory of Competition Among Pressure Groups for Political Influence," *The Quarterly Journal of Economics*, 98(3), 371–400.
- Berry, S.K. (1993). "Rent-seeking With Multiple Winners," *Public Choice*, 77, 437–443.
- Beviá, C. and L.C. Corchón (2010). "Peace Agreements Without Commitment," *Games and Economic Behavior*, 68(2), 469–487.
- Beviá, C. and L.C. Corchón (2013). "Endogenous Strength in Conflicts," *International Journal of Industrial Organization*, 31(3), 297–306.
- Beviá, C. and L.C. Corchón (2015). "Relative Difference Contest Success Function," *Theory and Decision*, 78(3), 377–398.
- Blavatskyy, P.-R. (2010). "Contest Success Function with the Possibility of a Draw: Axiomatization," *Journal of Mathematical Economics*, 46(2), 267–276.
- Brown, J. (2011). "Quitters Never Win: The (Adverse) Incentive Effects of Competing with Superstars," *Journal of Political Economy*, 119(5), 982–1013.
- Brown, A. and S.M. Chowdhury (2014). "The Hidden Perils of Affirmative Action: Sabotage in Handicap Contests," *University of East Anglia Working Paper No. 62*.
- Bulow, J., J. Geanakoplos and P. Klemperer (1985). "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*, 93(3), 488–511.
- Calsamiglia, C., J. Franke and P. Rey-Biel (2013). "The Incentive Effects of Affirmative Action in a Real-effort Tournament," *Journal of Public Economics*, 98, 15–31.
- Che, Y.-K. and I. Gale (2000). "Difference-form Contests and the Robustness of All-pay Auctions," *Games and Economic Behavior*, 30(1), 22–43.
- Chowdhury, S.M. and O. Gürler (2015). "Sabotage in Contests: A Survey," *Public Choice*, 164(1), 135–155.
- Chowdhury, S.M. and S.H. Kim (2014). "A Note on Multi-winner Contest Mechanisms," *Economics Letters*, 125, 357–359.
- Chowdhury, S.M. and R.M. Sheremeta (2011). "Multiple Equilibria in Tullock Contests," *Economics Letters*, 112(2), 216–219.
- Chowdhury, S.-M. and I. Topolyan (2015). "The Group All-pay Auction with Heterogeneous Impact Functions," *University of East Anglia Working Paper No. 69*.
- Chowdhury, S.-M. and I. Topolyan (2016). "The Attack-and-Defense Group Contests: Best Shot Versus Weakest Link," *Economic Inquiry*, 54(1), 548–557.
- Chowdhury, S.-M., D. Lee and R.M. Sheremeta (2013). "Top Guns May Not Fire: Best-shot Group Contests With Group-specific Public Good Prizes," *Journal of Economic Behavior and Organization*, 92, 94–103.
- Chowdhury, S.-M., D. Lee and I. Topolyan (2016). "The Max-Min Group Contest: Weakest-link (Group) All-pay Auction," *Southern Economic Journal*, 83(1), 105–125.
- Chung, T.-Y. (1996). "Rent-seeking Contest when the Prize Increases with Aggregate Efforts," *Public Choice*, 87(1), 55–66.

- Clark, D. and C. Riis (1996). "A Multi-winner Nested Rent-seeking Contest," *Public Choice*, 87, 177–184.
- Clark, D. and C. Riis (1998a). "Contest Success Functions: An Extension," *Economic Theory*, 11(1), 201–204.
- Clark, D. and C. Riis (1998b). "Competition Over More Than One Prize," *The American Economic Review*, 88(1), 276–289.
- Corchón, L.C. (2000). "On the Allocative Effects of Rent-seeking," *Journal of Public Economic Theory*, 2(4), 483–491.
- Corchón, L.C. (2007). "The Theory of Contests: A Survey," *Review of Economic Design*, 11(2), 69–100.
- Corchón, L.C. and M. Dahm (2010). "Foundations for Contest Success Functions," *Economic Theory*, 43(1), 81–98.
- Corchón, L.C. and M. Dahm (2011). "Welfare Maximizing Contest Success Functions When the Planner Cannot Commit," *Journal of Mathematical Economics*, 47(3), 309–317.
- Corchón, L.C. and A. Yıldızparlak (2013). "Give Peace a Chance: The Effect of Ownership and Asymmetric Information on Peace," *Journal of Economic Behavior & Organization*, 92, 116–126.
- Cornes, R. and R. Hartley (2003). "Risk Aversion, Heterogeneity and Contests," *Public Choice*, 117(1), 1–25.
- Cournot, A. A. (1838). *Recherches sur les Principes Mathématiques de la Théorie des Richesses*, Paris: Hachette.
- Cubel, M. and S. Sánchez-Pagés (2014). "An Axiomatization of Difference-form Contest Success Functions," *Working Paper*, University of Barcelona.
- Dagan, N. and O. Volij (1993). "The Bankruptcy Problem: A Cooperative Bargaining Approach," *Mathematical Social Sciences*, 26(3), 287–297.
- Dahm, M. and P. Esteve (2016). "Affirmative Action Through Extra Prizes," *Working Paper*.
- Dahm, M. and N. Porteiro (2008). "Biased Contests," *Public Choice*, 136(1/2), 55–67.
- Dechenaux, E., D. Kovenock and R.M. Sheremeta (2015). "A Survey of Experimental Research on Contests, All-pay Auctions and Tournaments," *Experimental Economics*, 18(4), 609–669.
- Dixit, A. (1987). "Strategic Behavior in Contests," *American Economic Review*, 77(5), 891–898.
- Einy, E., O. Haimanko and D. Moreno et al. (2015). "Equilibrium Existence in Tullock Contests with Incomplete Information," *Journal of Mathematical Economics*, 61, 241–245.
- Einy, E., O. Haimanko, D. Moreno and B. Shitovitz (2010). "On the Existence of Bayesian Cournot Equilibrium," *Games and Economic Behavior*, 68(1), 77–94.
- Epstein, G. and S. Nitzan (2006). "The Politics of Randomness," *Social Choice and Welfare*, 27(2), 423–433.
- Everhart, C. (2015). "Mixed Equilibria in Tullock Contests," *Economic Theory*, 60(1), 59–71.
- Falconieri, S., F. Palomino and J. Sákovics (2004). "Collective Versus Individual Sale of Television Rights in League Sports," *Journal of the European Economic Association*, 2(5), 833–862.
- Fang, H. (2002). "Lottery Versus All-pay Auction Models of Lobbying," *Public Choice*, 112(3), 351–371.
- Faria, J.R., F.G. Mixon, S.B. Caudill and S.J. Wineke (2014). "Two-dimensional Effort in Patent-race Games and Rent-seeking Contests: The Case of Telephony," *Games*, 5(2), 116–126.
- Franke, J. (2012a). "Affirmative Action in Contest Games," *European Journal of Political Economy*, 28(1), 105–118.
- Franke, J. (2012b). "The Incentive Effects of Levelling the Playing Field – An Empirical Analysis of Amateur Golf Tournaments," *Applied Economics*, 44(9), 1193–1200.
- Fu, Q. (2006). "Endogenous Timing of Contest with Asymmetric Information," *Public Choice*, 129(1), 1–23.
- Fu, Q., J. Lu and Y. Pan (2015). "Team Contests with Multiple Pairwise Battles," *The American Economic Review*, 105(7), 2120–2140.
- Fullerton, R.L. and R.P. McAfee (1999). "Auctioning Entry into Tournaments," *Journal of Political Economy*, 107(3), 573–605.
- Gelder, A., D. Kovenock and B. Roberson (2015). "All-pay Auctions with Ties," Chapman University, unpublished Manuscript, October 2015.
- Hillman, A. and E. Katz (1984). "Risk-averse Rent Seekers and the Social Cost of Monopoly Power," *The Economic Journal*, 94(373), 104–110.
- Hillman, A. and J. Riley (1989). "Politically Contestable Rents and Transfers," *Economics and Politics*, 1(1), 17–39.
- Hirschleifer, J. (1989). "Conflict and Rent-seeking Success Functions: Ratio vs. Difference Models of Relative Success," *Public Choice*, 63(2), 101–112.
- Hirschleifer, J. (1991). "The Technology of Conflict as an Economic Activity," *The American Economic Review*, 81(2), 130–134.
- Hwang, S.H. (2012). "Technology of Military Conflict, Military Spending, and War," *Journal of Public Economics*, 96(1–2), 226–236.
- Jackson, M. and M. Morelli (2011). "The Reasons for Wars: An Updated Survey," in C. Coyne (ed.), *Handbook on the Political Economy of War*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, 34.
- Jia, H. (2008). "A Stochastic Derivation of the Ratio Form of Contest Success Functions," *Public Choice*, 135(3), 125–130.
- Jia, H. (2012). "Contests with the Probability of a Draw: A Stochastic Foundation," *Economic Record*, 88(282), 391–401.
- Jia, H., S. Skaperdas and S. Vaidya (2013). "Contest Functions: Theoretical Foundations and Issues in Estimation," *International Journal of Industrial Organization*, 31(3), 211–222.

- Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47(2), 263–292.
- Katz, E., S. Nitzan and J. Rosenberg (1990). "Rent-seeking for Pure Public Goods," *Public Choice*, 65(1), 49–60.
- Kolmar, M. and H. Rommeswinkel (2013). "Contests with Group-specific Public Goods and Complementarities in Efforts," *Journal of Economic Behavior and Organization*, 89, 9–22.
- Konrad, K. (2009). "Strategy and Dynamics in Contests," New York: Oxford University Press.
- Konrad, K. (2012). "Dynamic Contests and the Discouragement Effect," *Revue d'Economie Politique*, 122(2), 233–256.
- Konrad, K. and M. Gradstein (1999). "Orchestrating Rent-seeking Contests," *The Economic Journal*, 109(458), 536–545.
- Konrad, K. and H. Schlesinger (1997). "Risk Aversion in Rent-seeking and Rent-augmenting Games," *The Economic Journal*, 107(445), 1671–1683.
- Klumpp, T. and M.K. Polborn (2006). "Primaries and the New Hampshire Effect," *Journal of Public Economics*, 90(6–7), 1073–1114.
- Krueger, A. (1974). "The Political Economy of the Rent-seeking Society," *American Economic Review*, 64(3), 291–303.
- Lee, D. (2012). "Weakest-link Contests with Group-specific Public Good Prizes," *European Journal of Political Economy*, 28(2), 238–248.
- Lee, S. and J.-H. Kang (1998). "Collective Contests with Externalities," *European Journal of Political Economy*, 14(4), 727–738.
- Leininger, W. (1993). "More Efficient Rent-seeking: A Münchhausen Solution," *Public Choice*, 75(1), 43–62.
- Levitt, S.D. (1994). "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House," *Journal of Political Economy*, 102(4), 777–798.
- Luo, Z. and X. Xie (2016). "A Theory of Rivalry with Endogenous Strength," mimeo.
- Milgrom, P. and R. Weber (1982). "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50(5), 1089–1122.
- Moldovanu, B. and A. Sela (2001). "The Optimal Allocation of Prizes in Contests," *The American Economic Review*, 91(3), 542–558.
- Möller, M. (2012). "Incentives versus Competitive Balance," *Economics Letters*, 117(2), 505–508.
- Morgan, J. (2003). "Sequential Contests," *Public Choice*, 116(1), 1–18.
- Münster, J. (2009). "Group Contest Success Functions," *Economic Theory*, 41(2), 345–357.
- Nash, J. (1950). "Equilibrium Points in N-person Games," *Proceedings of the National Academy of Sciences*, 36(1), 48–49.
- Nitzan, S. (1994). "Modelling Rent-seeking Contests," *European Journal of Political Economy*, 10(1), 41–60.
- Nitzan, S. and K. Ueda (2009). "Collective Contests for Commons and Club Goods," *Journal of Public Economics*, 93(1–2), 48–55.
- Nitzan, S. and K. Ueda (2011). "Prize Sharing in Collective Contests," *European Economic Review*, 55(5), 678–687.
- Nti, K.O. (1998). "Effort and Performance in Group Contests," *European Journal of Political Economy*, 14(4), 769–781.
- Olson M. (1985). *The Logic of Collective Action*, Cambridge, MA: Harvard University Press.
- Olszewski, W. and R. Siegel (2016a). "Large Contests," *Econometrica*, 84(2), 835–854.
- Olszewski, W. and R. Siegel (2016b). "Effort-maximizing Contests," *Working Paper*.
- Palomino, F. and J. Sákovics (2004). "Inter-league Competition for Talent vs. Competitive Balance," *International Journal of Industrial Organization*, 22(6), 783–797.
- Peeters, T. and S. Szymanski (2012). "Vertical Restraints in Soccer: Financial Fair Play and the English Premier League," *Working Paper* 2012028, University of Antwerp, Faculty of Applied Economics.
- Pérez-Castrillo, D. and T. Verdier (1992). "A General Analysis of Rent-seeking Games," *Public Choice*, 73(3), 335–350.
- Polischuk L.I. and A. Tonis (2013). "Endogenous Contest Success Functions: A Mechanism Design Approach," *Economic Theory*, 52(1), 271–297.
- Rosen, S. (1986). "Prizes and Incentives in Elimination Tournaments," *The American Economic Review*, 76(4), 701–715.
- Sela, A. (2011). "Best-of-three All-pay Auctions," *Economics Letters*, 112(1), 67–70.
- Serena, M. (2016). "Harnessing Beliefs to Stimulate Efforts," *SSRN Working Paper*, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2686543.
- Serena, M. (2017). "Quality Contests," *European Journal of Political Economy*, 46(C) 15–25.
- Siegel, R. (2009). "All-pay Contests," *Econometrica*, 77(1), 71–92.
- Siegel, R. (2010). "Asymmetric Contests with Conditional Investments," *The American Economic Review*, 100(5), 2230–2260.
- Sisak, D. (2009). "Multiple-prize Contests – The Optimal Allocation of Prizes," *Journal of Economic Surveys*, 23(1) 82–114.

146 *Handbook of game theory and industrial organization: applications*

- Skaperdas, S. (1991). "Conflict and attitudes toward risk," *The American Economic Review*, 81(2), 116–120.
- Skaperdas, S. (1996). "Contest Success Functions," *Economic Theory*, 7(2), 283–290.
- Skaperdas, S. and L. Gan (1995). "Risk Aversion in Contests," *The Economic Journal*, 105(431), 951–962.
- Skaperdas, S. and S. Vaidya (2012). "Persuasion as a Contest," *Economic Theory*, 51(2), 465–486.
- Szymanski, A. (2003). "The Economic Design of Sporting Contests," *Journal of Economic Literature*, 41(4), 1137–1187.
- Taylor, C.R. (1995). "Digging for Golden Carrots: An Analysis of Research Tournaments," *The American Economic Review*, 85(4), 872–890.
- Topolyan, I. (2014). "Rent-seeking for a Public Good with Additive Contributions," *Social Choice and Welfare*, 42(2), 465–476.
- Treich, C.R. (2010). "Risk-aversion and Prudence in Rent-seeking Games," *Public Choice*, 145(3), 339–349.
- Tullock, G. (1967). "The Welfare Cost of Tariffs, Monopolies and Theft," *Western Economic Journal*, 5(3), 224–232.
- Tullock, G. (1980). "Efficient Rent-seeking," in J.M. Buchanan, R.D. Tollison and G. Tullock (eds), *Towards a Theory of a Rent-Seeking Society*, College Station, TX: Texas A&M University Press, 97–112.
- Tullock, G. (2003). "The Origin Rent-seeking Concept," *International Journal of Business and Economics*, 2(1), 1–8.
- Vázquez-Sedano, A. (2014). "Sharing the Effort Costs in Group Contests," *SSRN Working Paper*: <http://ssrn.com/abstract=2439828>.
- Vesperoni, A. (2013). "A Contest Success Function for Rankings," mimeo, University of Siegen.
- Vrooman, J. (2012). "Two to Tango: Optimum Competitive Balance in Pro Sports Leagues," in P. Rodriguez, S. Kesenne and J. Garcia (eds), *The Econometrics of Sport*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Wang, Z. (2010). "The Optimal Accuracy Level in Asymmetric Contests," *The B.E. Journal of Theoretical Economics*, 10(1).
- Wärneryd, K. (2000). "In Defense of Lawyers: Moral Hazard as an Aid to Cooperation," *Games and Economic Behavior*, 33(1), 145–158.
- Wärneryd, K. (2003). "Information in Conflicts," *Journal of Economic Theory*, 110(1), 121–136.
- Yates, A. (2011). "Winner-pay Contests," *Public Choice*, 147(1), 93–106.
- Yıldızparlak, A. (2013). "A Contest Success Function for Ties with an Application to Soccer", *Working Paper*, Department of Economics, Universidad Carlos III de Madrid.



7. Endogenous timing in contests

*Magnus Hoffmann and Grégoire Rota-Graziosi**

1 INTRODUCTION

In game theory, there is a fundamental distinction between simultaneous-move games and sequential-move games. While in the former no knowledge of the strategies chosen by other players is available, in the latter the strategy of at least one player is known by other players. This distinction leads to two different equilibrium concepts, which are typically applied to particular games: the Cournot-Nash equilibrium (NE) in a simultaneous-move game and the subgame perfect Nash equilibrium (SPE), the so-called “Stackelberg equilibrium” (SE), in a sequential-move game.

It was von Stackelberg (1934) who first pointed out that, in sequential-move games, a first-mover advantage exists if two firms compete over quantities: each firm prefers to be the leader rather than the follower. Hence, as long as sequential-move games are based on the premise that the order of play (sequential) as well as the assignment of roles (leader and follower) is exogenously fixed, the question of the appropriateness of any particular order of moves emerges.¹ Consequently, starting in the early 1990s, several attempts have been made to endogenize the order of moves in various games,² with the seminal work of Hamilton and Slutsky (1990) as the most prominent.³

From a theoretical point of view, the purpose of these approaches was to address the uneasiness associated with the concept of SEs. From an empirical point of view, they clarified that distinct behavioral patterns could be rationalized by an endogenous timing approach. Overall, the associated findings made possible sharper predictions about (i) whether a sequential order of moves emerges and (ii) the particular assignment of roles.

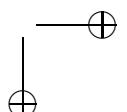
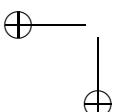
However, the majority of the above-mentioned literature assumes a monotone effect of players’ strategies on opponents’ marginal payoffs. If marginal payoffs are non-monotonic, with respect to the opponent’s strategy, the best responses also become non-monotonic. Whether players regard strategies as strategic substitutes (SS) or as strategic complements

* We would like to thank Philipp Denter, Georgios Katsenos, and Martin Kolmar for their insightful comments. Of course, any remaining errors are our own. This paper benefited from the financial support of the FERDI (Fondation pour les Études et Recherches sur le Développement International) and of the program “Investissement d’Avenir” (reference ANR-10-LABX-14-01) of the French government.

¹ Friedman (1983) even argued that the SE is too artificial a concept and, as such, cannot be observed in reality (p. 175).

² For example, Robson (1990) analyzed endogenous sequencing in a Bertrand competition framework. Albæk (1990) endogenized the order of moves in a Cournot competition framework with cost uncertainty. Deneckere and Kovenock (1992) examined the case of price leadership in the presence of capacity constraints. See also Mailath (1993) and Daughety and Reinganum (1994).

³ The specific role of commitment has been emphasized by Schelling (1960). A player’s ability to commit may improve his outcome at the equilibrium significantly. Indeed, moving first is often considered as taking an advantage. However, depending on the nature of the game, moving second also provides a significant advantage, called the “second-mover advantage.” While moving first is an unconditional commitment in Schelling’s terminology, moving second is a conditional one, which consists of playing the best response with respect to the action of the leader.



(SC) then depends on the opponents' strategy.⁴ While non-monotonic best responses remain an exception in the industrial organization (IO) literature,⁵ contests are rarely super- or submodular.⁶ Hence, this property of contests involves a more thorough analysis and prohibits the application of tools, such as lattice theory, which are frequently used when facing monotone comparative statics (see, for instance, Vives, 2001, p. 16 *et seq.*).

This chapter surveys the literature on endogenous timing in contests.⁷ In Section 2, we will present the structure of a two-player contest with simultaneous as well as sequential moves when the prize is fixed. This will give us the basis for the analysis of the endogenous timing game (ETG) in Section 3. In Section 4, we will present ubiquitous contests with effort-dependent prizes, while Section 5 presents the structure of two-player contests with simultaneous as well as sequential moves against the background of effort-dependent prizes. The associated ETG is illustrated in Section 6. Section 7 surveys the literature on sequential play and ETG in contests that goes beyond the hitherto discussed. Section 8 concludes.

2 SIMULTANEOUS-MOVE AND SEQUENTIAL-MOVE CONTESTS WITH A FIXED PRIZE

Dixit (1987) was the first to analyze the behavior of players in a sequential-move contest with a fixed prize. In particular, he showed that, due to the players' non-monotonic best responses, no *a priori* prediction on whether players have an incentive to overexert or to underexert effort, compared to the simultaneous-move Nash equilibrium (NE), is possible. Moreover, he showed that there is a correlation between a player's probability to win at the NE and his strategic incentives as a first-mover in a sequential-move contest, if the contest-success function (CSF), which maps efforts into probabilities of winning, is either of the logit type or probit type.⁸ In particular, he finds that the favorite, i.e., the player whose probability of winning at the NE exceeds one-half, has an incentive to overexert effort, while the underdog, i.e., the player whose win probability at the NE is below one-half, has an incentive to underexert effort in the Stackelberg equilibrium (SE). If either player has a win probability of one-half at the NE, then he shows that the first-mover has no local incentive to deviate from his level of effort. We will now present this point in more detail, using the example of a logit-type CSF.

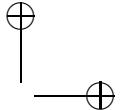
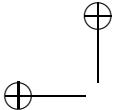
⁴ A player regards strategies as strategic complements (substitutes), if an increase of the opponent's strategy increases (decreases) his marginal payoff (see Bulow, Geanakoplos, and Klemperer, 1985b).

⁵ For instance Bulow, Geanakoplos, and Klemperer (1985a) showed that non-monotonicities emerge in a Cournot market, if the price elasticity of demand is constant. Amir, Amir, and Jin (2000) presented a model of R&D decisions in a Cournot oligopoly in which, at a pre-play stage, each firm invests in order to reduce the marginal costs of production. It is assumed that knowledge spillovers emerge and it is shown that, depending on the rate of those spillovers, non-monotonic best responses emerge at the R&D stage. See Hoffmann and Rota-Graziosi (2014) for a detailed survey on the literature of games with non-monotonic payoffs and games with non-monotonic marginal payoffs.

⁶ This particular property of contests is an artifact of the assumed contest-success function. See, for instance, Konrad (2009, p. 67 *et seq.*).

⁷ Other reviews on contests that preceded the present one are *inter alia* Corchón (2007) and Garfinkel and Skaperdas (2007). Particular attention to sequential-move contests, as well as to endogenous timing, were paid by Nitzan (1994, pp. 51–52), Hirshleifer (1995b, pp. 168–169), Konrad (2009, pp. 67–71), and Dechenaux, Kovenock and Sheremeta (2014, pp. 630–631).

⁸ For a more detailed review of the different CSFs frequently used in contests, see Hirshleifer (1989), Corchón (2007), Konrad (2009, pp. 23–53) and Corchón and Serena (Chapter 6 in this volume).



Consider a situation in which each of two players exerts effort $x_i \in \mathbb{R}_+$ in order to win a prize. Player i 's valuation of the prize is $V_i > 0$ and his probability of winning is given by a logit-type CSF:

$$p^i(\mathbf{x}) = \frac{f^i(x_i)}{f^i(x_i) + f^j(x_j)}, \quad (7.1)$$

for $\mathbf{x} > 0$, with $\mathbf{x} = (x_1, x_2)$, $i, j \in \{1, 2\}$ and $i \neq j$.⁹ Moreover, we assume that $f_i^i(x_i) > 0 \geq f_{ii}^i(x_i)$ and $f^i(x_i) \geq 0$ hold for $x_i > 0$.¹⁰ The direct costs of effort are $C^i(x_i)$, with $C_i^i(x_i) > 0$ and $C_{ii}^i(x_i) \geq 0$. Then the payoff function of player i becomes:

$$\Pi^i(\mathbf{x}) = p^i(\mathbf{x}) V_i - C^i(x_i). \quad (7.2)$$

By exerting effort, player i induces a spillover, which we will call ‘‘CSF-dependent spillover effect’’: *ceteris paribus*, the win probability of the opponent decreases. This becomes clear if we take a look at the cross-effect on the payoff function:

$$\Pi_j^i(\mathbf{x}) = p_j^i(\mathbf{x}) V_i = -\frac{f_j^j(x_j) f^i(x_i) V_i}{(f^i(x_i) + f^j(x_j))^2} < 0. \quad (7.3)$$

Using the terminology of Eaton (2004), eq. (7.3) shows that we have a game of plain substitutes (PS), as opposed to a game with plain complements (PC), which emerges if $\Pi_j^i(\mathbf{x}) > 0$ for both players. Thus, if a player is able to credibly commit to a level of effort, he will do this in such a manner that the effort of the other player will be reduced. Whether this means to over- or underexert effort compared to the NE depends on the slope of the rival's best response function at the NE.

If we assume an interior solution, then player i 's best response function is implicitly given by the first-order condition (FOC):

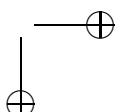
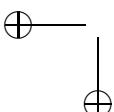
$$\Pi_i^i(\mathbf{x}) = 0 \Leftrightarrow p_i^i(\mathbf{x}) V_i - C_i^i(x_i) = 0. \quad (7.4)$$

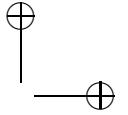
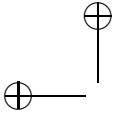
The sign of the slope of the best response function at a point in the strategy space is solely determined by the cross-effect on the marginal payoff function, since the denominator of eq. (7.5) is unambiguously negative:

$$\frac{dx_i}{dx_j} = -\frac{\Pi_{ij}^i(\mathbf{x})}{\Pi_{ii}^i(\mathbf{x})} = -\frac{p_{ij}^i(\mathbf{x}) V_i}{p_{ii}^i(\mathbf{x}) V_i - C_{ii}^i(x_i)}. \quad (7.5)$$

⁹ $p^i(\mathbf{x})$ can also be interpreted as the *share* of the prize V_i that player i wins with a probability of one (as, for instance, in Grossman and Kim, 1995). Both interpretations are consistent, if the players are assumed to be risk neutral. In the remainder of this chapter, we will adopt the probabilistic interpretation.

¹⁰ Index i (j) is the derivative with respect to player i 's (j 's) effort. To avoid repetition, we use $i, j = 1, 2$ and $i \neq j$ when it is obvious.





The nominator's sign in eq. (7.5), however, depends exclusively on the cross-effect on the marginal win probability:

$$p_{ij}^i(\mathbf{x}) = (f^i(x_i) - f^j(x_j)) \frac{f_i^i(x_i)f_j^j(x_j)}{(f^i(x_i) + f^j(x_j))^3} \left\{ \begin{array}{l} \geq \\ \leq \end{array} \right\} 0 \Leftrightarrow p^i(\mathbf{x}) \left\{ \begin{array}{l} \geq \\ \leq \end{array} \right\} \frac{1}{2}. \quad (7.6)$$

Evaluating the vector of efforts at the NE, $\mathbf{x}^N = (x_i^N, x_j^N)$, and using the terminology of Bulow et al. (1985b), one finds that the underdog regards efforts as strategic substitutes (SS), while the favorite regards them as strategic complements (SC). Thus, a Stackelberg leader facing an underdog as a second-mover will always overexert effort compared to the NE level, in order to reduce the competitor's effort. Facing a favorite, however, an underdog-leader will underexert effort, in order to reduce the favorite's effort. If both players have equal win probability, then neither of them regards efforts as SS or SC, and both regard efforts as strategically independent (SI). In this case, no first-mover has an incentive to deviate locally from the NE level of effort. Example 1 represents a simplified version of Dixit (1987):

Example 1 (Dixit, 1987) Suppose two players compete over a prize with valuation $V_i > 0$. Player i 's cost function is given by $C^i(x_i) = c_i x_i$ and the probability of winning is given by a symmetric lottery CSF (a logit-type CSF with $f(x) = x$). Then we find that $x_i^N = c_j V_i^2 V_j (c_j V_i + c_i V_j)^{(-2)}$, so that $\Pi_{12}^1(\mathbf{x}^N) \left\{ \begin{array}{l} \geq \\ \leq \end{array} \right\} \Pi_{12}^2(\mathbf{x}^N) \Leftrightarrow p^1(\mathbf{x}) \left\{ \begin{array}{l} \geq \\ \leq \end{array} \right\} p^2(\mathbf{x}) \Leftrightarrow \frac{c_2}{V_2} \left\{ \begin{array}{l} \geq \\ \leq \end{array} \right\} \frac{c_1}{V_1}$. Hence, for the case $c_1 V_2 = c_2 V_1$, we get $p^1(\mathbf{x}^N) = p^2(\mathbf{x}^N)$, while for $c_1 V_2 < c_2 V_1$ player 1 (2) becomes the favorite (underdog) of the game. The slope of player i 's best response function at the NE then becomes $\frac{dx_i}{dx_j} \Big|_{x_i \rightarrow x_i^N, x_j \rightarrow x_j^N} = \frac{1}{2} \left(\frac{c_j V_i}{V_j c_i} - 1 \right)$.

The two cases are represented in Figures 7.1 and 7.2. The bold graphs represent the players' best response functions, and the grey surface represents the strategy profiles that dominate the

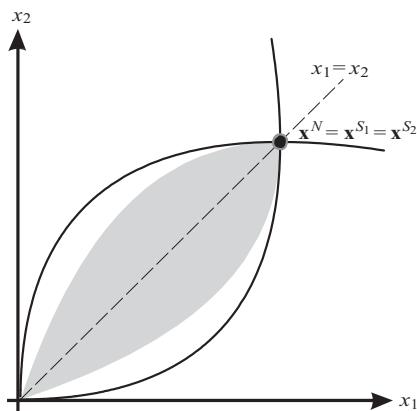
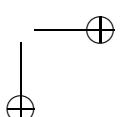
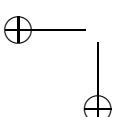


Figure 7.1 Example 1: Dixit (1987), symmetric case



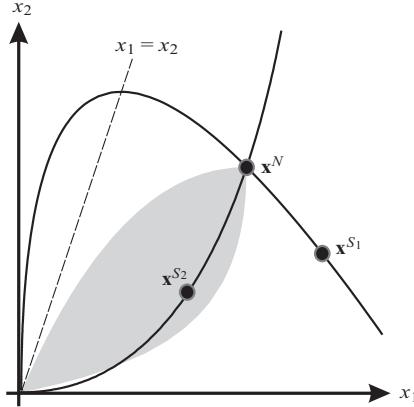


Figure 7.2 Dixit (1987), asymmetric case

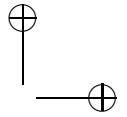
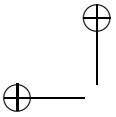
NE in a Pareto sense.¹¹ In the symmetric case, the strategy profile at the NE equals the ones at the Stackelberg equilibria (SE). x_i^L hereby represents the leader's effort, while x_j^F represents the follower's effort. Consequently, $\mathbf{x}^{S_i} = (x_i^L, x_j^F)$ represents the vector of efforts at the sub-game perfect equilibrium (SPE) of a sequential-move game, in which player i leads and player j follows. In the asymmetric case, the leader-underdog (player 2) reduces his effort, while the leader-favorite (player 1) increases his effort in comparison to the NE level, i.e., $x_2^L < x_2^N$ and $x_1^L > x_1^N$. As a reaction, the follower always decreases his own effort compared to the NE level, i.e., $x_1^F < x_1^N$. Hence, \mathbf{x}^{S_1} lies to the south-east, while \mathbf{x}^{S_2} lies to the south-west of \mathbf{x}^N .

Note that, if the payoff function is given by eq. (7.2), then $x_i^N > 0$, although a level of $x_i(\tilde{x}_i)$ always exists, such that the best response of player j becomes zero for all values of x_i beyond \tilde{x}_i : $BR^j(x_i) = 0, \forall x_i \geq \tilde{x}_i$. Whether it is possible that, with symmetric players, the first mover preempts any effort of his followers was analyzed by Leininger and Yang (1994). They find that if two symmetric players compete over a prize using a Tullock CSF (a logit-type CSF with $f(x) = x^r$) with $r \geq 2$, then the Stackelberg leader deters any effort by the follower in the SPE of the sequential move game. Moreover, they demonstrate that sequential moves over an infinite period of time lead to implicit collusion, i.e., players choose a tit-for-tat strategy that precludes escalation in the choice of effort.¹²

In the case of asymmetric players, deterring effort is not as restricted, a fact that was first shown by Leininger (1993) and Linster (1993). They showed in the case of a lottery CSF, that deterrence emerges in the SPE of a fixed-prize sequential contest, if the Stackelberg leader (say i) is sufficiently strong, i.e., if $p^i(\mathbf{x}^N) \gg \frac{1}{2}$.

¹¹ Here, we assume that the relevant payoffs are exclusively those of the two contestants. In a broader sense, one might also incorporate the contest designer's payoff, which might very well be an increasing function of both contestants' effort, such as in sports competitions (see, for example, Szymanski, 2003, and Runkel, 2011).

¹² See also Pérez-Castrillo and Verdier (1992), who showed that, in the case of one Stackelberg leader and multiple followers, it is beneficial for the leader to implement a level of effort such that all subsequent players, who choose their effort simultaneously and independently in the subsequent stage, choose zero effort. This holds if the CSF is of the Tullock type and has increasing returns to scale ($r > 1$). A closer examination of the case with multiple followers and leaders can be found in Glazer and Hassin (2000).



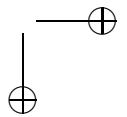
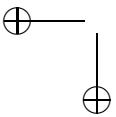
These results have been utilized in the literature on the endogenous enforcement of property rights. In Grossman and Kim (1995) and Grossman (2001), for example, it is shown how, in a state of anarchy, initial claims to property can be converted into effective property rights in a sequential-move game. Example 2 is a minimal working example of Grossman and Kim (1995):

Example 2 (Grossman and Kim, 1995) Suppose a defender (player 1) is endowed with a rent of value V_1 . This endowment is a claim to property, which is subject to appropriation. The appropriator's (player 2's) valuation of the rent is V_2 . The probability of successfully defending the initial claims to property is given by a lottery CSF and the marginal costs of player i are c_i . The payoffs are then given by eq. (7.2). Now, suppose that payoffs are common knowledge and that the defender moves first in order to be able to enforce the initial claims to property. One finds at the SE that $p^1(\mathbf{x}^{S_1}) = 1$, with $x_1^L = \frac{c_2}{4V_2} \left(\frac{V_1}{c_1} \right)^2$ and $x_2^F = 0$ if, and only if, player 1 is sufficiently relatively strong, i.e., if $\frac{c_1}{V_1} \leq \frac{c_2}{2V_2}$, so that $p^1(\mathbf{x}^N) \geq \frac{2}{3}$ (cf. Example 1).

Variations of this theme can be found, for instance, in Anderton, Anderton, and Carter (1999), Kolmar (2008) and Hoffmann (2010). Anderton et al. (1999) analyzed a two-player model of production, appropriation and trade, in which each player's fortification effort is exerted before the competitor's appropriation effort and the probability of winning is given by an asymmetric lottery CSF. Deterrence emerges in the SPE of the game, if the effectiveness of the appropriation effort is below the fortification effort. Hoffmann (2010) used a similar framework, in which players choose their level of fortification and appropriation efforts, as well as the amount of goods they want to exchange. In his model, the anticipation of potential appropriation forces the agents to engage in trade. In the SPE of the game the resulting post-trade allocation of consumption goods becomes uncontested if defense and appropriation are sufficiently unequally effective. Kolmar (2008) analyzed a model in which the prize is endogenous, in the sense that the defender can determine the value of the prize independently of the effort decision via a production function. Labor and effort are linked by a budget constraint that is assumed to be non-binding in optimum. Kolmar (2008) compared different timing scenarios. While, in the first stage, the production decision is made, the subsequent effort decisions are either simultaneous or sequential, with either the defender (as in Grossman and Kim, 1995) or the appropriator acting as a Stackelberg leader. It is shown that deterrence only emerges in the SPE of the game if the contest is sufficiently asymmetric and the defender acts as a Stackelberg leader.

In general, applications of sequential-move contests can be found in various areas, such as open conflict (Hirshleifer, 1988, 1995a), civil war (Azam, 1995), resource conflicts (Hotte, 2001) and litigation (Hirshleifer and Osborne, 2001). An overview of real-world examples for sequential move contests can be found in Morgan (2003).

Before we turn to the ETG in the fixed-prize scenario, we will briefly survey the literature on experiments that examine sequential move contests. Shogren and Baik (1992) carried out an experiment to assess the effect of sequential moves in contests, given an asymmetric lottery contest. The experiment only partly supported the predictions of Dixit (1987). In particular, leader-favorites did not always overexert effort when compared to the NE level of effort and the follower-underdogs frequently exerted effort that was above their best response. As mentioned above, Leininger and Yang (1994) showed that deterrence emerges in the SPE



of a symmetric and sequential move contest if the CSF is of the Tullock type with $r \geq 2$. However, as Weimann, Yang, and Vogt (2000) showed in an experiment, if $r > 2$, then the predicted SPE never emerges: either first-movers show cooperative behavior by bidding low and second-movers exploit this attempt by choosing their best response, or the first-movers bid high in order to preempt the effort exerted by the follower and the followers choose to punish that attempt, which leads to a negative payoff for both players. Vogt, Weimann, and Yang (2002) designed an experiment based on the open-end bidding game of Leininger and Yang (1994) in a Tullock contest framework with $r > 2$. The findings confirm that players learn how to avoid an escalation of bids in the suggested framework. In particular, average rent-seeking expenditures frequently fall below 10 percent of the prize within four rounds. Fonseca (2009) experimentally studied players' behavior in a two-player contest within a simultaneous-move as well as in a sequential-move framework. Contestants faced a symmetric lottery CSF in one treatment and an asymmetric one in a different treatment. In the latter, asymmetries were so large that deterrence emerged in the predicted SPE. In the laboratory, the author finds that, in the symmetric treatment (asymmetric treatment with the favorite leading), the leaders' bids were, on average, above (below) predictions. Moreover, in the asymmetric treatment, even if bids were large enough to prevent any positive bid by the follower-underdog, then the second-mover responded by choosing bids far above zero.

3 ENDOGENOUS TIMING WITH A FIXED PRIZE

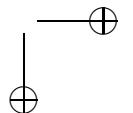
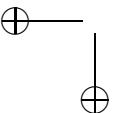
Up until now, we have assumed that the order of moves is exogenously fixed, a fact that had been criticized in the IO literature by various authors (see, for instance, Albæk, 1990 and Shapiro, 1989, p. 390). As a consequence, the order of moves became the center of attention in various applications in the following years, of which the work of Hamilton and Slutsky (1990) is the most prominent.

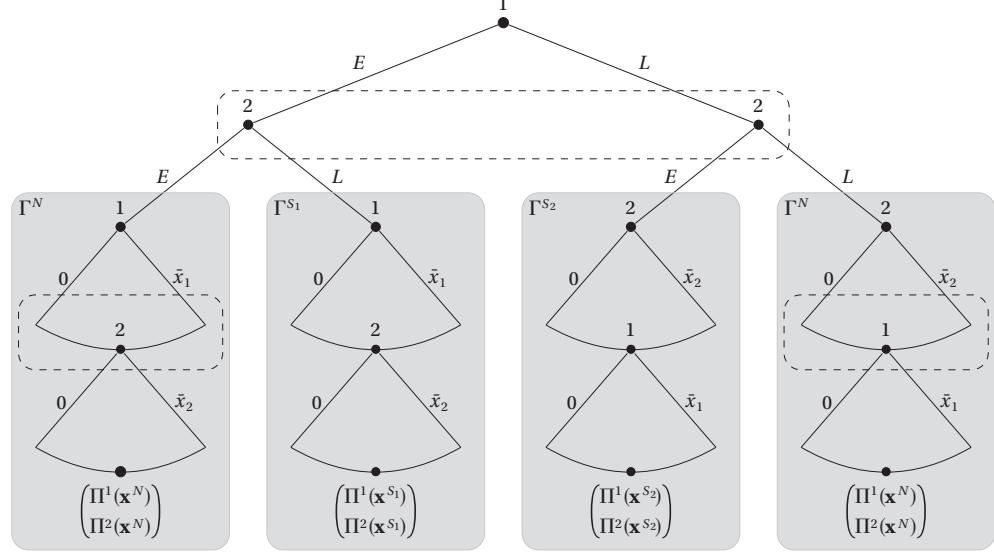
The first to undertake the task of endogenizing the sequence of moves in a contest framework were, independently, Baik and Shogren (1992) and Leininger (1993). Both papers assumed two agents with payoff functions identical to eq. (7.2).¹³ They endogenized the order of moves by using a framework that is similar to the *extended game with observable delay* in Hamilton and Slutsky (1990). Here, both players decide in a pre-play stage whether they want to exert effort as *soon* as, or as *late* as, possible. The decision of either player is then publicly announced. In the subsequent contest subgame, both players choose their effort according to their timing decisions in the pre-play stage.¹⁴ Hence, the basic game consists of three different constituent games: Γ^{S_1} (Γ^{S_2}) if player 1 (2) chooses E in the pre-play stage and player 2 (1) chooses L , and Γ^N if both players choose the same strategy in the pre-play stage. The extended game ($\tilde{\Gamma}$) is represented in extensive form in Figure 7.3.¹⁵

¹³ While Baik and Shogren (1992) assume a general logit-type or probit-type CSF with asymmetries towards the impact function ($f^1(x) \neq f^2(x)$) and a common valued prize, Leininger (1993) assumes an asymmetric lottery CSF with asymmetric valuations of the prize. Both make the assumption of unit direct costs of effort.

¹⁴ It is worth mentioning that none of the players can gain by renegeing on his pre-play decision. See Hamilton and Slutsky (1990, p. 32).

¹⁵ By assumption, the strategy space of each player in any constituent game is unbounded above. However, one can think of \bar{x}_i as the level of effort such that $\Pi^i(x_i, 0) < 0, \forall x_i > \bar{x}_i$. Therefore for all $x_i > \bar{x}_i$, \bar{x}_i strictly dominates x_i . Additionally, since $\Pi_j^i(x) < 0$, this inequality also holds for any $x_j > 0$. Hence, after the elimination of those strictly dominated strategies, the strategy space of player i becomes $[0, \bar{x}_i]$.



Figure 7.3 The extensive form of $\tilde{\Gamma}$ with the SPE in each basic game

The key observation of Baik and Shogren (1992) and Leininger (1993) was that, in the case of asymmetric payoffs, a unique SPE of the extended game exists. In this SPE, the sequential order of moves emerges in the basic game, with the underdog leading and the favorite following. In order to explain the behavior in the pre-play stage, we need to compare the different payoffs in the different subgames.

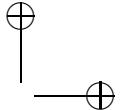
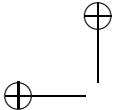
Player i has a first-mover incentive, if his Stackelberg leader payoff exceeds his payoff at the NE, i.e., if $\Pi^i(\mathbf{x}^{S_i}) > \Pi^i(\mathbf{x}^N)$. It is straightforward to show that this condition holds as long as the cross-partial derivative of the competitor's marginal payoff function is non-zero, i.e., $\Pi_{ij}^j(\mathbf{x}^N) \neq 0$ (cf. eq. 7.5). With symmetric players, we find that no player has a first-mover incentive so that $\mathbf{x}^N = \mathbf{x}^{S_1} = \mathbf{x}^{S_2}$. Then, since $\Pi_{ij}^j(\mathbf{x}) = 0$, all possible sub-games yield the same equilibrium payoffs for both players. Hence, there are multiple NEs in the pre-play stage that are all payoff-equivalent.

Player i has a second-mover incentive, if his Stackelberg follower payoff exceeds his NE payoff, i.e., if $\Pi^i(\mathbf{x}^{S_j}) > \Pi^i(\mathbf{x}^N)$. A second-mover incentive only exists if player i regards efforts as SC: a Stackelberg leader j will reduce effort compared to x_j^N , in order to reduce the best response of player i , $BR^i(x_j)$:

$$\Pi^i(x_i^N, x_j^N) < \Pi^i(x_i^N, x_j^L) \leq \max_{x_i} \Pi^i(x_i, x_j^L) \equiv \Pi^i(x_i^F, x_j^L), \quad (7.7)$$

where the first inequality stems from the fact that we have a game of PS. Player i has no second-mover incentive if he regards efforts as SS: a Stackelberg leader j will increase effort compared to x_j^N , in order to reduce the best response of player i :

$$\Pi^i(x_i^F, x_j^L) < \Pi^i(x_i^F, x_j^N) \leq \max_{x_i} \Pi^i(x_i, x_j^N) \equiv \Pi^i(x_i^N, x_j^N). \quad (7.8)$$



Supposing that $p^1(\mathbf{x}^N) > \frac{1}{2}$, then both players have a first-mover incentive (since $\Pi_{ij}^i(\mathbf{x}^N) \neq 0$, see eq. 7.5 and 7.6), but only the favorite (player 1) has a second-mover incentive (since $\Pi_{12}^1(\mathbf{x}^N) > 0 > \Pi_{12}^2(\mathbf{x}^N)$).¹⁶ Thus, we get

$$\Pi^i(\mathbf{x}^{S_i}) > \Pi^i(\mathbf{x}^N), \quad \Pi^1(\mathbf{x}^N) < \Pi^1(\mathbf{x}^{S_2}), \quad \Pi^2(\mathbf{x}^{S_1}) < \Pi^2(\mathbf{x}^N). \quad (7.9)$$

Consequently, the underdog has a dominant strategy in the pre-play stage, since both the Stackelberg leader payoff and the NE payoff exceed the payoff as a Stackelberg follower. Therefore his dominant strategy is to move *early*. This can best be seen by the reduced form representation of $\tilde{\Gamma}$ in Table (7.1). Given this, the favorite's best response is to choose *late*, since he has a second-mover incentive. The SPE of the contest subgame therefore shows a sequential order of moves, with the underdog moving first. Hence, the key determinant of endogenous timing in the fixed-prize framework is the win probability at the NE.

This result is interesting on many levels. First, there is either a unique equilibrium to the timing game, or there are multiple equilibria that all lead to the same equilibrium efforts in the contest subgame. Second, if players' payoffs are asymmetric, then the unique equilibrium of the timing game shows sequential moves. This not only reduces the uneasiness associated with sequential moves in general games (see Friedman, 1983, p. 175), but also within contests in particular (see Tullock, 1980, p. 107 and Tullock, 1985, p. 260). Third, the identity of the Stackelberg leader as well as follower coincides with the identity of the favorite and underdog, respectively. Fourth, and most importantly, there is a welfare effect. If spending effort is viewed as a social cost, then the SPE of the total game Pareto-dominates the SE with the favorite leading as well as the NE. Thus, endogenizing the order of moves unambiguously reduces rent dissipation and leads to a second-best efficient outcome.

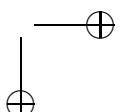
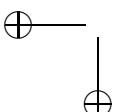
This result shows that, while the assumption of simultaneous moves in contests is appropriate if players are symmetric, sequential moves can easily be justified if there are asymmetries between players in the above-described underlying game. Also, in the absence of a contest designer, who is able to determine the exact order of moves, the assumption of simultaneous moves in asymmetric contests becomes harder to justify in the light of these results.

The endogenous timing approach has been applied to several contests' frameworks with fixed prizes, such as environmental disputes (Baik, 1994, and Shogren and Hurley, 1997)

Table 7.1 The reduced form of $\tilde{\Gamma}$

		Player 2	
		Early	Late
Player 1	Early	$\Pi^1(\mathbf{x}^N), \Pi^2(\mathbf{x}^N)$	$\Pi^1(\mathbf{x}^{S_1}), \Pi^2(\mathbf{x}^{S_2})$
	Late	$\Pi^1(\mathbf{x}^{S_2}), \Pi^2(\mathbf{x}^{S_2})$	$\Pi^1(\mathbf{x}^N), \Pi^2(\mathbf{x}^N)$

¹⁶ Note that the same player may have a first-mover and second-mover incentive in the same game. In contrast, first-mover and second-mover advantages are mutually exclusive.



or conflicts over property rights (Kolmar, 2008). Experiments partially confirmed the results of Baik and Shogren (1992) and Leininger (1993). In particular, Shogren and Hurley (1997) experimentally studied behavior in a two-player contest, given an asymmetric lottery CSF. The results of that study show that underdogs only wished to lead 35–58 percent of the time, while the favorites only wanted to follow 55–77 percent of the time. The experiment also showed that leader-underdogs frequently overexerted effort and leader-favorites underexerted effort.

Baik et al. (1999), in a similar framework, experimentally examined an asymmetric contest with endogenous timing, in which the contestants had two days to think about their strategies. The strategy set consisted of a timing and a corresponding effort decision. They found that timing decisions and bidding behavior were close to the theoretical predictions. Notably, 82 percent of the time, effort levels and timing decisions followed the theoretical predictions.

4 ENDOGENOUS PRIZE CONTESTS

In a strict sense, the above-derived results are only applicable in two-player contests in which the contestants fight over a fixed prize. However, since the late 1980s, there is a steadily growing literature on contests that allow for an effort-dependent prize.¹⁷ For this strand of the literature, the above-derived results are not applicable, which will be shown in Section 6. As will become apparent, in the presence of an endogenous prize, the set of strategic incentives and, therefore, the timing approaches that can be justified by an ETG, is richer than in the fixed-prize scenario. Below, we will survey the literature on endogenous prize contests, in order to give the reader an overview of the multiplicity of the applications of such contests in IO and beyond the IO literature. In the subsequent sections, we will then analyze whether the assumed order of moves in the literature (mostly simultaneous moves) can be motivated by the assumption of an underlying ETG that explicitly allows for effort-dependent, i.e., endogenous prizes.

Conflict models Prizes are endogenous if the marginal costs of exerting effort are solely represented by the marginal reduction in the value of the prize. This kind of effort-dependent prize emerges in a class of models that has been labeled *conflict models*, as opposed to *rent-seeking* models in which the direct costs of effort are assumed to be exogenous.¹⁸ In this class of models, an agent usually faces a trade-off between two strategies, which have been coined *butter and guns*: (i) a productive use of an inalienable resource (e.g., labor) that will create value to a common pool good, or (ii) a non-productive use of the resource (exerting effort) in order to increase the share of the common pool good one is able to consume. Thus, the reduction of the prize resembles the opportunity costs of exerting effort in this economy. Hirshleifer (1991b) examined the effect of relative strength on equilibrium payoffs. Skaperdas (1992) and Beviá and Córchón (2010) investigated under which conditions open conflict can be avoided in the absence of formal property rights, while Anbarci, Skaperdas,

¹⁷ Konrad (2009) showed that the literature reveals various modeling approaches to endogenize prizes in contests: the value of the prize may be seen as a strategy of a contest designer (see Appelbaum and Katz, 1987, Moldovanu and Sela, 2001 or Che and Gale, 2003); as a direct strategy of a player (see Gradstein, 1993, Konrad, 2002 or Hoffmann, 2010); or as a function of the players' efforts. Henceforth, we will refer to the term "endogenous prize" in the latter sense.

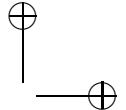
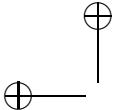
¹⁸ The difference between rent-seeking and conflict models, or, to put it differently, partial and total equilibrium models, was first proposed by Neary (1997).

and Syropoulos (2002) compared different bargaining solutions in which the threat-point is contingent on the players' investments in effort. Grossman (2001) showed the impact of initial claims to property in conflict models. Skaperdas and Syropoulos (2002) and Anderton and Carter (2008) analyzed the impact of insecure property rights on trade, and Hodler and Yektaş (2012) investigated the impact of asymmetric information on equilibrium behavior.

War and other forms of attrition In military conflicts, the destruction of the competitor's valuable resources is something that is usually put up with by the adversaries. In terms of economic modeling, this kind of environment is often represented by a prize-production function that decreases in military expenditures as, for example, in Shaffer (2006). In Grossman and Kim (1995), the competitors' efforts are aimed at changing the property rights of a valuable good (the prize). However, if the relative strength of one competitor suffices to appropriate a share of the good, then the value of this good decreases. In this model, the prize is a non-continuous and non-increasing function of the relative effort exerted by both agents (see also Garfinkel and Skaperdas, 2000). Cai (2003) and Smith et al. (2014) used an effort-dependent prize, which continuously decreases in the combatants' military expenditures.

Research & development Baye and Hoppe (2003) showed that certain types of research tournaments are strategically equivalent to an endogenous-prize Tullock contest, in which the prize is monotonically increasing in the players' effort (as, for example, in Chung, 1996). In Kaplan, Luski, and Wettstein (2003), firms choose a time of innovation in an all-pay auction. The novelty of this approach is that the winning firm's reward decreases in the firm's strategy, so that earlier innovation not only increases the probability of winning but also the value of the reward. Kaplan et al. (2002) complemented the previous paper by assuming asymmetric information regarding the type of opponent. Zhou (2006), in a tournament setting, compared the equilibrium R&D spending, if the prize is fixed, to a setting in which the prize increases in the winning firm's spending. Clark and Riis (2007) showed in which manner an endogenous-prize Tullock contest can be utilized, in order to find a suitable partner for an R&D firm.

Promotional effort In promotional competition, firms spend effort in order to increase their market share, which is contingent on a firm's relative expenditure (see Schmalensee, 1976). The advertising expenditures are sunk costs so that promotional competition has an all-pay structure. Barros and Sørgard (2001) analyzed in a contest framework, the welfare effects of mergers when firms spend effort on promotion. The latter activity not only increases the market share of a firm (determined by a Tullock CSF), but also the demand for the products in the market *per se*, since advertising increases consumer product awareness. Chioveanu (2008) presented a model of persuasive advertising in which the number of consumers who react to advertising positively is a function of total advertising expenditures in the market. The share of consumers that become loyal to a specific brand is then determined via a *market-sharing function*, which is contingent on advertising expenditures and that exhibits all the properties of a general CSF. Effort-dependent prizes also emerged in Haan and Moraga-González (2011), who provided a tournament model of attention-seeking in general markets and in De Frutos, Ornaghi, and Siotis (2013) concerning the pharmaceutical industry. Ridlon (2013) analyzed retailer competition, in which advertising has market size as well as market share effects; Espinosa and Mariel (2001) offer a dynamic model.

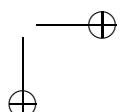
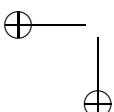


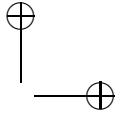
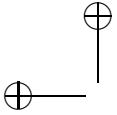
Reimbursement schemes Generally, reimbursement schemes in contests can be interpreted as endogenous-prize games: for example, in Matros and Armanios (2009), the winner or the loser of a probabilistic rent-seeking contest is partially reimbursed for his effort by a contest designer. Thus, the winner's and loser's prize monotonically increases in own effort. If the winner is fully reimbursed, then multiple equilibria exist, which has been analyzed by Cohen and Sela (2005) and Matros (2012). Comparisons of alternative reimbursement schemes can be found in Hurley (1998), Lim and Shogren (2004) and Park (2010). Cohen and Shavit (2012) experimentally investigated the difference in equilibrium effort between contests with and without reimbursement.

Litigation In litigation, the decision to go to court is heavily influenced by the fee-shifting rule of the justice system, which determines whether the winner (or loser) has to compensate the legal fees of his opponent (see Spier, 2007). In this sense, all costs are borne by the players (litigants). For example, Farmer and Pecorino (1999) showed that a trial under the “British rule,” in which the loser has to reimburse the costs of the winner, can be interpreted as a Tullock contest, in which the prize is contingent on the expenditures for legal representation. Baye, Kovenock, and de Vries (2005) compared several fee-shifting rules in an all-pay auction setting. With the exception of the “American rule,” all legal systems share the common feature that spillovers emerge contingent on the players’ ranking of efforts. Structurally equivalent settings in a contest framework yield winner and loser prizes, that are effort dependent (see Chowdhury and Sheremeta, 2011). Baik and Shogren (1994) examined the legal dispute between a citizens’ group and a firm over environmental regulations with and without fee-shifting rules (see also Gong and McAfee, 2000 for a general setting). Bernardo, Talley, and Welch (2000) analyzed the effect of different forms of compensation schemes in legal disputes in an all-pay auction setting. Wärneryd (2000) presented a two-stage model of delegation, in which the optimal contract that a litigant should offer to his lawyer turns the latter’s objective function into one in which the prize is monotonically decreasing in the lawyer’s effort.

Promotion tournaments and relative-performance reward schemes Gershkov, Li, and Schweinzer (2009) analyzed a model of team production, in which the total prize is a monotonically increasing function of the team members’ efforts. The winner and loser prizes are shares of the total prize and both shares are part of a take-it-or-leave-it offer to one of the players, which is made by a team member at an early stage of the game. The model shows that efficient production within a team is implementable if both the winner and the loser prize are effort dependent. Cohen, Kaplan, and Sela (2008) analyzed reward schemes in which the awarded prize is effort dependent and carefully chosen by a contest designer. Amegashie (2001) examined the case in which an applicant’s expected salary depends on the same person’s performance at the job interview, so that the probability of getting the job as well as the salary depend positively on the exerted effort.

Other-regarding preferences Baye, Kovenock, and de Vries (2012) presented a general model of rank-order spillovers in all-pay auctions. In one of their applications, they adopt the idea of inequality aversion, as originated by Fehr and Schmidt (1999), with respect to the invested effort. In case a player’s effort is larger (smaller) than that of his competitor, he experiences an additional disutility proportional to the difference in effort. Thus, a player who exerts larger (smaller) effort in comparison to his opponent receives a prize that, *ceteris*





paribus, decreases (increases) in his own effort. Hoffmann and Kolmar (2013) showed that this approach is structurally equivalent to a probabilistic contest in which players are, *ex ante*, inequality averse with respect to their income, or, in other words, in which the benchmark for income comparisons is the expected rather than the *ex post* income.

Fund-raising lotteries Lotteries have a long history of financing public goods (see Lange, List, and Price, 2007). Morgan (2000) compared lotteries to alternative methods of fundraising for public goods (for example, voluntary contributions). The probability of winning the lottery prize is determined by a Tullock-lottery CSF and the sum of total wagers represents the government's revenues from which a share will be used to finance a public good, while another represents the awarded prize. In a particular setting (the pari-mutuel raffle), the prize is assumed to be revenue-dependent, i.e., the prize-production function is a linear function of the sum of all the wagers. In a similar framework, Gregor (2012) compared linear and non-linear prize-production functions in lotteries. Dale (2004) compared fixed-prize and revenue-dependent prize lotteries that finance a public good by means of an experiment.

5 SIMULTANEOUS-MOVE AND SEQUENTIAL-MOVE CONTESTS WITH AN ENDOGENOUS PRIZE

We will now turn to the analysis of simultaneous-move and sequential-move contests in the presence of an endogenous prize. As it turns out, endogenizing the prize has a substantial effect on the players' strategic incentives and, thus, on the SPE of the extended game. The reason for this change in the strategic incentives is due to the additional externality stemming from the effort-dependent prize. In a fixed-prize framework, the effort exerted by one player produces a single kind of spillover: *ceteris paribus*, the win probability of the opponent decreases. This "CSF-dependent spillover effect" is accompanied by another in the presence of effort-dependent prizes: a "prize-dependent spillover effect" always exists, i.e., the effort of at least one player alters his competitors' payoff via that player's effort effect on the prize.

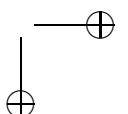
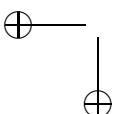
Suppose that the prize faced by player i is given by $V^i(\mathbf{x})$, with $V^i(\mathbf{0}) > 0$. Then player i 's payoff function becomes

$$\Pi^i(\mathbf{x}) = p^i(\mathbf{x}) V^i(\mathbf{x}) - C^i(x_i). \quad (7.10)$$

Hoffmann and Rota-Graziosi (2012) showed in the case of an effort-dependent prize with a common value, i.e., $V^1(\mathbf{x}) = V^2(\mathbf{x}) = V(\mathbf{x})$, that the strategic incentives in a two-player contest are heavily influenced by the structure of the prize-production function, in particular the cross-partial derivative of the prize-production function. Hoffmann and Rota-Graziosi (2012) assumed that both players' effect on the prize is non-positive. Together with some mild assumptions on the degree of heterogeneity of players, this leads to (i) a unique NE in the contest subgame and (ii) a game of plain substitutes, since $\Pi_j^i(\mathbf{x}) < 0, \forall \mathbf{x} > \mathbf{0}$.

In comparison to eq. (7.5), the sign of the slope of player i 's best response function at the unique NE now becomes

$$\Pi_{ij}^i(\mathbf{x}^N) = p^i(\mathbf{x}^N) V_{ij}(\mathbf{x}^N) + \Omega^i(\mathbf{x}^N), \quad (7.11)$$





with

$$\Omega^i(\mathbf{x}^N) = \frac{p_i^i(\mathbf{x}^N) C_j^j(x_j^N)}{p_j^j(\mathbf{x}^N)} - \frac{p_j^j(\mathbf{x}^N) C_i^i(x_i^N)}{p_i^i(\mathbf{x}^N)}.$$

Since $\Omega^1(\mathbf{x})$ and $\Omega^2(\mathbf{x})$ are symmetric functions, one finds that the following relationship holds at the NE:

$$\Pi_{12}^1(\mathbf{x}^N) + \Pi_{12}^2(\mathbf{x}^N) = V_{12}(\mathbf{x}^N). \quad (7.12)$$

If $V_{12}(\mathbf{x}^N) = 0$, then either both players' incentives are directly opposed ($\Pi_{ij}^i(\mathbf{x}^N) > 0 > \Pi_{ij}^j(\mathbf{x}^N)$) or are aligned and equal to zero ($\Pi_{ij}^i(\mathbf{x}^N) = 0 = \Pi_{ij}^j(\mathbf{x}^N)$). In the latter case, there may be a favorite and an underdog or each competitor's win probability is one-half.

Example 3 (Beviá and Corchón, 2010) Beviá and Corchón (2010) presented a conflict model with a Tullock CSF and a common valued prize $V(\mathbf{x}) = \omega - k_1x_1 - k_2x_2$, with $\omega > 0$, $k_1 = k_2$, so that $V_{12}(\mathbf{x}) = 0$, $\Omega^i(\mathbf{x}) = 0$ and $\Pi_{12}^1(\mathbf{x}^N) = \Pi_{12}^2(\mathbf{x}^N) = 0$. This resembles a symmetric conflict model so that $p^1(\mathbf{x}^N) = p^2(\mathbf{x}^N)$. Implementing an asymmetry between players such that $k_1 > k_2$ still leads to $\Pi_{12}^1(\mathbf{x}^N) = \Pi_{12}^2(\mathbf{x}^N) = 0$, but now $p^1(\mathbf{x}^N) < p^2(\mathbf{x}^N)$, since the opportunity costs of effort are larger for player 1. Figure 7.4 represents the asymmetric case.

In Example 3, we learn that, even when both players regard efforts as SI, a favorite and an underdog may exist. Hence, all three basic games show the exact same efforts in equilibrium, i.e., $\mathbf{x}^N = \mathbf{x}^{S_1} = \mathbf{x}^{S_2}$. The reason for this is the following. First, any asymmetry between players clearly triggers the relative NE effort, and therefore, the win probabilities. Second, due to the fact that the cross-partial derivative of the prize-production function is zero, the leader cannot manipulate the follower's marginal calculation.

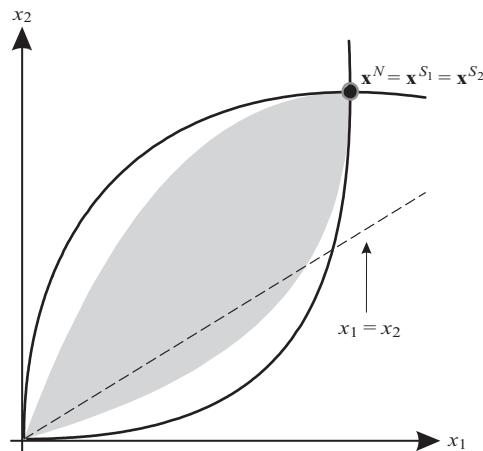


Figure 7.4 Example 3: Beviá and Corchón (2010)





Next, we find that players may have opposite strategic incentives, even if no favorite and no underdog exist:

Example 4 (Shaffer, 2006) In Shaffer (2006), we find a rent-seeking model with a lottery CSF and a common valued prize $V(\mathbf{x}) = 1 - k_1 x_1 - k_2 x_2$, with $k_1 = k_2$. Moreover, direct costs of effort exist, with $C^i(x_i) = c_i x_i$ and $c_1 = c_2 = 1$. This resembles a symmetric rent-seeking model with common values, so that $p^1(\mathbf{x}^N) = p^2(\mathbf{x}^N)$ and, since $V_{ij}(\mathbf{x}) = 0$, $\Pi_{12}^1(\mathbf{x}^N) = \Pi_{12}^2(\mathbf{x}^N) = 0$ at the unique NE. Implementing two types of asymmetries between the players, such that $c_1 = 1 < c_2 = \frac{3}{2}$ and $k_1 = 2 > k_2 = 1$ leads to strategies that are directly opposed ($\Pi_{12}^1(\mathbf{x}^N) > 0 > \Pi_{12}^2(\mathbf{x}^N)$), although still $p^1(\mathbf{x}^N) = p^2(\mathbf{x}^N)$. Figure 7.5 represents the asymmetric case.

In Example 4, we learn that, while the asymmetry regarding the marginal impact on the prize ($k_1 \neq k_2$) triggers the NE level of effort, the asymmetry regarding the marginal costs ($c_1 \neq c_2$) has an impact on the NE level as well as on the strategic incentives at the NE, since $\Omega^1(\mathbf{x}^N) > 0 > \Omega^2(\mathbf{x}^N)$ (cf. eq. 7.11).

However, more can be inferred from eq. (7.11). It also follows that both players may regard efforts as SS (only compatible with $V_{ij}(\mathbf{x}^N) < 0$), or that both players regard efforts as SC (only compatible with $V_{ij}(\mathbf{x}^N) > 0$), which cannot emerge in a fixed-prize contest. The latter case, for example, emerged in Skaperdas (1992), in which $V_{ij}(\mathbf{x}) > 0$ and direct costs of effort are zero.

Example 5 (Skaperdas, 1992) In Skaperdas (1992), two players each possess one unit of an inalienable resource, which can be used to produce two kinds of inputs, y and x . y will be used in the joint production of the prize and x represents the players' effort. Thus, the individual budget constraint leads to $x_i = 1 - y_i$. It is assumed that $V_{ij}(\mathbf{x}) > 0$, so that $\Pi_{ij}^i(\mathbf{x}^N) = p^i(\mathbf{x}^N) V_{ij}(\mathbf{x}^N) > 0$, since $x_i^N \in (0, 1)$. An asymmetric case is represented in Figure 7.6.

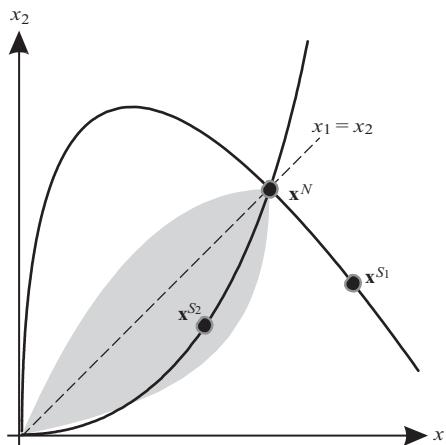


Figure 7.5 Example 4: Shaffer (2006)



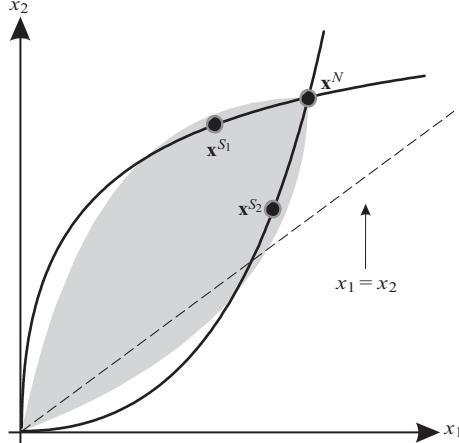


Figure 7.6 Example 5: Skaperdas (1992)

Thus far, we have excluded the literature on asymmetric valuation of the prize (as in Baye et al. 2005). These cases have been analyzed by Hoffmann (2015). Suppose that the payoff function is given by

$$\Pi^i(\mathbf{x}) = p^i(\mathbf{x}) V^i(\mathbf{x}) - C^i(x_i), \quad (7.13)$$

and that we have a symmetric game, i.e., $V_1^1(\mathbf{x}) = V_2^2(\mathbf{x})$ and $V_1^2(\mathbf{x}) = V_2^1(\mathbf{x})$. Then, Hoffmann (2015) found that

$$\Pi_{ij}^i(\mathbf{x}^N) = p_i^i(\mathbf{x}^N) (V_j^i(\mathbf{x}^N) - V_i^i(\mathbf{x}^N)). \quad (7.14)$$

In this case, the players' incentives are aligned. They may regard efforts as SS, SC or SI, contingent on the difference of the marginal impact on the prize at the NE. An example of this kind of game is presented in Alexeev and Leitzel (1996):

Example 6 (Alexeev and Leitzel, 1996) Alexeev and Leitzel (1996) presented a rent-seeking model with a lottery CSF, in which $n \geq 2$ players compete over a rent of value $\omega > 0$ and marginal costs are constant and equal to one. It is assumed that the prize is decreasing, i.e., its value is reduced by the sum of outlays of all losing parties, so that $V^i(\mathbf{x}) = \omega - x_j$ in a two-player setting. Thus, $V_j^i(\mathbf{x}) = -1$ and $V_i^i(\mathbf{x}) = 0$, so that $\Pi_{ij}^i(\mathbf{x}^N) = -p_i^i(\mathbf{x}^N) < 0$ at the unique symmetric NE and both players regard efforts as SC. This case is represented in Figure 7.7.¹⁹

Comparing the results in Sections 2 and 5, we learn that, while the win probability is crucial for the determination of the strategic incentives in the fixed-prize scenario (cf. eq. 7.6), it is

¹⁹ Note that, for an interior solution of both sequential-move games, we need to assume that marginal costs are sufficiently high, which turns out to be a value considerably larger than one.

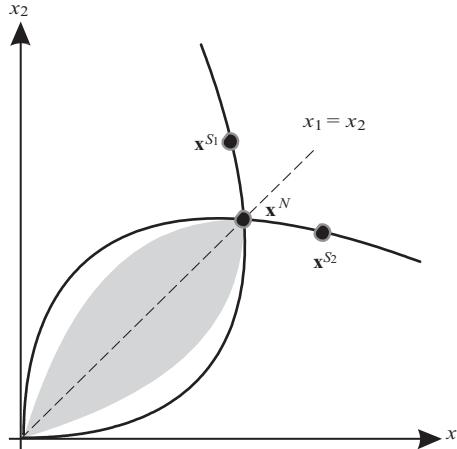


Figure 7.7 Example 6: Alexeev and Leitzel (1996)

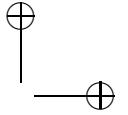
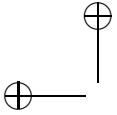
the prize-production function (for example, in conflict games with common valuation of the prize), or the prize-production function in collaboration with the win probability (for example, in rent-seeking games with common valuation of the prize) that determines the strategic incentives in the endogenous-prize scenario (cf. eq. 7.11 and eq. 7.14).

6 ENDOGENOUS TIMING GAMES WITH AN ENDOGENOUS PRIZE

Given the results of the previous section, we can now turn to the corresponding endogenous timing game. The results of the previous sections show that, as in the fixed-prize scenario, either no player has an incentive to deviate from the NE level of effort, or one player overexerts effort as a leader in the appertaining SE, while the competitor underexerts effort. However, the primitives of the model that drive these results are different when comparing to the fixed-prize framework.

The first case (with $\Pi_{ij}^i(\mathbf{x}^N) = 0 = \Pi_{ij}^j(\mathbf{x}^N)$), for example, emerges in conflict games independent of the players' win probability if $V_{ij}(\mathbf{x}^N) = 0$, cf. Example 3 and Figure 7.4. Hence, even if players are asymmetric, no first-mover incentive exists for either of them, and thus the strategies chosen in the SPE of each constituent game are identical ($\mathbf{x}^N = \mathbf{x}^{S_1} = \mathbf{x}^{S_2}$). Consequently, players are indifferent between their strategies in the pre-play stage of $\tilde{\Gamma}$ (cf. Table 7.1). This result partly reinforces the statement by Tullock (1980, 1985), who interpreted contests primarily as simultaneous-move games. Furthermore, it strengthens the modeling assumptions regarding the sequence of moves of, for instance, Garfinkel and Skaperdas (2000), and Beviá and Córchón (2010).

The second case (with $\Pi_{ij}^i(\mathbf{x}^N) > 0 > \Pi_{ij}^j(\mathbf{x}^N)$), for example, emerges in rent-seeking games and then depends on both the win probability and the structure of the prize-production function (cf. Example 4 and Figure 7.5). Hence, even if players are symmetric, first-mover incentives may exist for either of them, so that $\mathbf{x}^{S_i} \neq \mathbf{x}^N$. Then, a unique SPE of $\tilde{\Gamma}$ exists, with sequential moves in the basic game. However, in contrast to the fixed-prize scenario,



the identity of the first-mover remains unclear, since there is no correlation between win probability and strategic incentives.

Moreover, we saw in endogenous-prize contests that both players may regard efforts as SS or as SC. In the latter case, both players have a first-mover as well as a second-mover incentive, since $\Pi_{ij}^i(\mathbf{x}^N) \geq \Pi_{ij}^j(\mathbf{x}^N) > 0$. This case naturally emerges, for instance, in conflict games when $V_{ij}(\mathbf{x}^N) > 0$, cf. Example 5 and Figure 7.6. Given this case, both players prefer both SEs over the NE in the basic game that, when turning to the decisions in the pre-play stage presented in Table 7.1, yields two pure strategy equilibria: (E, L) and (L, E) . Accordingly, the SPE of the ETG yields sequential play in the basic game. Therefore, the ETG is not able to support the timing approaches utilized in, for example, Hirshleifer (1991a).

If both players regard efforts as SS ($\Pi_{ij}^i(\mathbf{x}^N) \leq \Pi_{ij}^j(\mathbf{x}^N) < 0$), then both players have a first-mover incentive but no second-mover incentive. This case emerges, for example, in symmetric contests if the difference of the marginal impact of the competitor's and the player's own effort is negative (cf. eq. 7.14, Example 6 and Figure 7.7). Hence, we find that $\Pi^i(\mathbf{x}^{S_i}) > \Pi^i(\mathbf{x}^N) > \Pi^i(\mathbf{x}^{S_j})$ and moving early becomes a dominant strategy in the pre-play stage for both players (cf. Table 7.1). Hence, the SPE of the ETG shows simultaneous moves by both players. This result strengthens the modeling assumptions regarding the sequence of moves of, for instance, Alexeev and Leitzel (1996) and Matros and Armanios (2009).²⁰

7 SPECIAL TOPICS

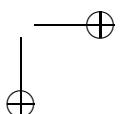
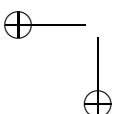
In the following sections, we will discuss some topics in the area of endogenous timing that go beyond the previous study.

7.1 Information Asymmetry

In this section, we focus on information asymmetry within the context of endogenous timing contests. There are mainly two situations that have been studied in the literature: (i) each opponent's costs of effort and abilities are private information; (ii) the prize's true value for the contestants is private information. These two sources of information asymmetry can be analyzed through the same approach. Indeed, keeping the costs of effort private would correspond to keeping the value of the prize private. Hurley (1998) and Fey (2006) establish that information asymmetry on the cost of effort of each contestant reduces equilibrium effort expenditures in a simultaneous rent-seeking game.

Linster (1993) introduces information asymmetry in a sequential rent-seeking game. He considers that the leader is unsure of the follower's type, regarding the valuation of the prize. He concludes that, with and without information asymmetry, total rent-seeking expenditures are higher when the leader is the contestant who values the prize more. Wärneryd (2003) extends this by assuming that the value of the prize is distributed according to a cumulative distribution function. The author shows that, under some conditions on the shape of the CSF, the uninformed agent is the favorite at the simultaneous-move NE.

²⁰ In Matros and Armanios (2009), both players regard efforts as SS, if the marginal compensation of effort associated with the loser prize is larger than it is for the winning prize.



Following these works, Fu (2006) goes a step further by introducing information asymmetry into an endogenous timing rent-seeking game. The value of the prize is fixed and identical for both agents. However, this is only known by one agent: the informed party. If the likelihood of a prize of low value is sufficiently low, the SPE of the ETG would correspond to the Stackelberg outcome, in which the informed party will move second. Otherwise, players will move simultaneously. Thus, informational advantage translates into a weakly dominant strategy for the informed agent to move later (second-mover advantage).²¹ Following Morgan (2003), Fu (2006) illustrates his results with the dates of US National Presidential Conventions, emphasizing that the incumbent or the representative of the incumbent's party (the informed party) always declared his candidacy after his opponent during the period of 1948–2004.

The information asymmetry may concern other aspects of the contest. For instance, Baik and Kim (2014) consider contests with delegation in which delegation contracts are private information. Thus, the delegates, when they choose their respective effort, do not know the contract and the value of the prize of their respective opponent. At the NE, the higher-valuation player offers his delegates less compensation when contracts are private information, while the lower-valuation player offers more. The initial gap between agents in terms of prize valuation is then reduced through delegation and information asymmetry.

In Morath and Münster (2013), contestants choose to pay to know the value of the prize. The authors consider several cases: the decision to acquire information is observable while the information on the value remains private; this decision is private; the acquired information is common knowledge. Depending on the cost of information, one or both agents may invest in information acquisition. Such a move can be viewed as an investment, which is decided before the contest takes place.

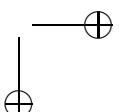
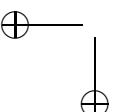
Baumann, Denter, and Friehe (2013) formalize a private precaution against crime. In this contest, the defender has private information on the value of the prize and decides to make his defense effort observable or not. This choice is equivalent to the defender choosing a simultaneous-move contest and a sequential-move one, in which he is the leader. In this setup, the ETG is partial since only the defender has the ability to choose his timing. When both players know the value of the prize, the defender moves first and deters his potential attacker. In contrast, when the value of the prize is private information for the defender, the latter may choose to play simultaneously or, equivalently, not to disclose his defense effort at the equilibrium. The authors emphasize that, in their setup, incomplete information reduces effort expenditures.

Finally, Denter and Sisak (2015) show that deterrence in conflicts is less likely in sequential-move games, under the assumption of incomplete information regarding the follower's type. In particular, a leader who would deter effort of the opponent in the complete information scenario, will refrain from doing so under even the slightest bit of uncertainty.

7.2 Endogenous Timing Games as a Specific Commitment

Moving first or second is a type of commitment and the ETG can be viewed as a specific commitment game. Schelling (1960) highlights the role of commitments in improving players' outcomes. He extends the approach of von Stackelberg (1934), who emphasized

²¹ Actually, the informed agent is indifferent between following or playing the simultaneous-move game.



the advantage of moving first (leading) in a Cournot duopoly. Schelling (1960) defines the notion of “pure commitment” (p. 188) as a definite commitment to a pure strategy “equivalent to ‘first move’ in a two-person, two-move game, in which one would otherwise have to move second.” Schelling (1960) also considers conditional commitment, i.e., a commitment to a pure strategy that is contingent on the strategies played by others. Moreover, he also considers fractional threat, i.e., a commitment to mixed strategies. Several authors have recently formalized some intuitions of Schelling (1960). Rosenthal (1991) and Van Damme and Hurkens (1996) define the notion of commitment robust equilibrium based on moving first. Their respective definitions differ slightly, in particular with respect to commitment in mixed strategies. However, the commitment robustness of the Nash equilibrium of a given game is directly related to the SPE of the ETG version of this game: the NE of a static game is commitment robust if it is also the SPE of the corresponding ETG.

Renou (2009) proposes a broad definition of a commitment game, which is “a two-stage game in which the game played in the second-stage is endogenously determined by the commitment in the first stage” (p. 489). In this regard, the ETG is a particular kind of commitment game. In Schelling’s terminology, leading or playing x_i^L corresponds to a pure unconditional commitment, while following (playing $BR^j(x_i^L)$) is a conditional commitment. Bade, Haerlinger, and Renou (2009) generalize the ETG and establish a theorem that reduces the analysis of successful commitments to the study of simple commitments, in which one player engages himself in a unique action at the first stage of the game and the other player commits to a subset of actions that contains his best response to the commitment of his opponent. Finally, Kalai et al. (2010) consider conditional commitments: each player chooses not only his strategy but also the set of available actions. The authors establish a commitment folk theorem, where all feasible payoffs are reachable at the equilibria of the commitment game.

The commitment approach allows us to encompass and compare the ETG with other forms of commitment games. For example, we can compare the approach of the SPE of ETG with the approach of accumulation games, proposed by Saloner (1987) and applied to fixed-prize rent-seeking games by Yildirim (2005). In this game, each player can add to his previous efforts after observing his rival’s effort. Commitment is then gradual, but still irreversible and perfectly observable. The author establishes that, at the SPE, the underdog never leads. This result contrasts with the SPE of the ETG with a fixed prize. The intuition is that the flexibility of exerting effort several times induces the favorite to act more aggressively and the underdog to prefer to act simultaneously. A generalization of Yildirim (2005) to contests with an endogenous prize can be found in Hoffmann (2015).

7.3 Endogenous Timing Games as Part of Sophisticated Commitments

We can also consider a more complex framework in which commitments take a more sophisticated form. In addition to moving early or late, other commitment technologies may be available to the players: for instance, introducing an investment, considering a delegation stage or determining the sharing rule among the members of each group before the ETG takes place, which is equivalent to considering a combination of several commitments of different types. The properties of SC, SS, PC, and PS would then be modified with respect to the initial ETG. The resulting rankings of the equilibrium effort and the existence of first-mover (second-mover) advantages (incentives) would then also be affected.

Following Baik and Kim (1997), who introduce strategic delegation in contests, Baik and Lee (2013) extend the ETG approach by considering delegation. In an initial stage, players simultaneously choose their respective delegate and, more precisely, the level of the compensation paid to the delegate if the latter wins the prize. Subsequently, the delegates play the ETG and receive higher contingent compensation than in the simultaneous-move game. As in the standard ETG, the underdog or, equivalently the delegate with the lower contingent compensation, chooses to lead in order to soften the competition, while the favorite prefers to follow. Baik and Lee (2013) emphasize that players prefer the simultaneous-move framework, while delegates prefer the ETG, which allows them to obtain higher compensation. As a consequence, delegation reverses the usual result that moving sequentially in the presence of strategic complementarity improves agents' payoff.

Münster (2007) introduces an investment stage before a perfectly discriminating contest (all-pay auction) takes place. At this preliminary stage, each player invests in their abilities: the invested amount reduces the player's cost of effort in the contest stage. While efforts are always chosen simultaneously, investment decisions may be simultaneous, sequential, or even the result of an endogenous timing game. If investments are unobservable, then efforts and investment are taken simultaneously and, at the unique equilibrium of the contest, rent dissipation is complete. The expected bids are higher with investments than without. If investments are observable, then players act strategically at the two stages of the game. Rent dissipation may become incomplete when the timing of investment decisions is endogenous. In this case, and under some conditions regarding the investment technology, only one agent, namely the first-mover, would invest, thus deterring the other from doing the same. Equilibrium efforts are then lower.

As delegation or investment, internal sharing rules in a group, which participates in a contest, are also a strategic device. In Nitzan (1991a), a group of players competes against another group of players in a probabilistic rent-seeking contest for a private good in which the group's effort is simply the sum of the efforts of all group members. If a group wins the contest, an internal sharing rule (which is contingent on the relative effort of each player) decides how the private good will be split up between all members of the group. Thus, exerting effort increases not only the probability of winning the between-group contest, but also the value of the prize, i.e., the share of the private good for a player in the winning group. Applications for this kind of framework can be found in politics, territorial disputes, and even sports (see Konrad, 2009, p. 124). A comparison of different sharing rules is provided by Nitzan (1991b), Noh (1999) and Sun and Ng (1999). The endogenous determination of sharing rules under the assumption of perfect information is discussed in Lee (1995), under the assumption of asymmetric information in Baik and Lee (2007), Nitzan and Ueda (2011) and Baik and Lee (2012). The influence of risk aversion on individual equilibrium effort is investigated in Konrad and Schlesinger (1997). Hausken (2000) examines different forms of within-group cooperation. Noh (1999) considers internal conflict among groups that are competing for a prize. The intra-group sharing rule may be decided simultaneously or sequentially. This choice impacts the total group effort in the appropriation activity. The author establishes that a sequential decision of sharing rules would induce the egalitarian rule, which reduces appropriation efforts and maximizes welfare. The egalitarian rule seems to play the role of a "pacifier" inside each group and then in society in general.

Some of these sophisticated commitments may be reduced to a simpler model of ETG. For instance, the game with a preliminary investment stage, which was proposed by Münster (2007), is also equivalent to an ETG with an endogenous prize.

7.4 Imperfectly Observable Commitment

The previous analysis assumes implicitly that the commitment of the leader is perfectly and freely observable by the follower. If this observation is not perfect, or if it imposes a cost to the follower, then the value of commitment, mainly understood as moving first (leading), may disappear and consequently the SPE(s) of the ETG may be modified. No one has completely analyzed these imperfections in endogenous timing in rent-seeking games, yet. However, several authors have analyzed the consequences of imperfect observation of commitment in a Stackelberg game and some of them derive their results as sequential-move contests.

Bagwell (1995) analyzed a “noisy-leader game” – a Stackelberg game in which the follower observes a signal of the first-mover’s choice. This author concludes that the value of commitment disappears, or equivalently, that all SPEs of the sequential-move game with a noise on the leader’s actions are payoff-equivalent to the NE of the static game. The potential follower selects his best-response action for all signal values. In other words, he ignores the signal’s value and behaves as if he were playing in the static game. The potential leader is then constrained to playing the static game too, thus abandoning his first-mover advantage. However, Van Damme and Hurkens (1997) restore the value of commitment by considering mixed-strategy equilibria and proposing a general theory of equilibrium selection based on the notion of risk dominance developed by Harsanyi and Selten (1988).

Morgan and Várdy (2007) address the issue of imperfect observation of commitment by assuming that the follower has to pay a small cost to observe the leader’s effort. Observation of the leader’s action is then endogenously chosen by the follower himself, who accepts to pay or not to observe. Similarly to Bagwell (1995), Morgan and Várdy (2013) establish that commitment has no value: all SPEs of the sequential-move contest with observation costs correspond to the NE of the static game. The intuition relies on the best responses of the two players. The concavity of the leader’s payoff function is a necessary condition for an interior solution of the Stackelberg game: the best response of the leader to any strategy of the follower will be a pure strategy since the action space is continuous. At the Stackelberg equilibrium, the follower can then perfectly anticipate the leader’s decision and has no incentive to pay to observe it, even if this cost is arbitrarily small. In turn, this cancels out any strategic impact the leader could have on the follower. Each player then plays their static game’s best response. By playing a pure strategy the leader destroys the follower’s incentive to pay to observe his action in equilibrium. As long as the action space is continuous, the strict concavity of the leader’s payoff function eliminates any mixed strategy. Morgan and Várdy (2013) generalize this analysis by considering imperfect competition in general. They emphasize the “fragility” of commitment as soon as observation is endogenized.

In contrast to the previous approach, Fu, Gürler, and Münster (2013) consider communication as a strategic device, which complements or substitutes efforts in contests. Sending a message is presented as a kind of commitment, which increases the incentive to win the contest. The authors assume that each player faces a cost when he sends a message of confidence but later loses the contest. Such a cost is more of a reputational cost than a pure communication cost, as in Morgan and Várdy (2007). At the equilibrium, the favorite reinforces his position by signaling his strength and decreasing the underdog’s effort. Similarly, Baumann et al. (2013) assumed that one player, i.e., the defender, can choose to be observed or not by the other player, i.e., the attacker. By making his defense effort observable, the defender chooses to play as the leader. This increases his payoff at the SPE in such a way that the defender is even

ready to pay to be observed. This result holds as long as the value of the prize, i.e., the type or the wealth of the defender, is common value. If this information is private, then the defender may have interest to hide his defense effort, preferring to play simultaneously.

While the follower has to pay to observe in Morgan and Várdy (2007), he may be paid by the leader to observe in Baumann et al. (2013). Thus, if we consider the ETG with observation cost as a commitment game with a sophisticated commitment technology, we may reconcile previous opposing approaches by considering a game, in which, at a preliminary stage, each player may choose to commit to pay or not to observe (to be observed) if he wants to follow (lead). The commitment to pay this cost may be self-enforcing if the player has a second (first)-mover advantage. In such a game, commitment may keep some value, especially if one or both players have a second-mover advantage.²²

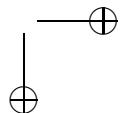
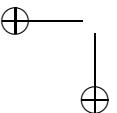
8 CONCLUSION

This chapter reviewed the ongoing literature regarding the timing of efforts in contests. We highlighted the determining role of the prize-production function on players' strategic incentives, on the shape of their respective reaction functions, and finally on the nature of the SPE of the ETG. In the presence of a fixed prize the underdog would lead at the SPE of the ETG. This equilibrium Pareto-dominates the NE of the corresponding static game: endogenizing the order of moves unambiguously reduces rent dissipation in this case. With an effort-dependent prize, the story becomes more subtle. Each player's effort now has two effects: one on the opponent strategy (the "CSF-dependent spillover effect") and another on the prize itself (the "prize-dependent spillover effect"). The prize-production function is then also determining the nature of the SPE of the ETG, which can correspond to a simultaneous-move Nash equilibrium or to one sequential-move and subgame perfect Nash equilibrium. Since both players may regard effort as SS or SC, the identity of the leader remains uncertain. We illustrated this review with several applications of ETG with fixed or endogenous prizes in the contest literature: research and development, promotion, litigation, reimbursement schemes, conflicts, war, attrition, environmental disputes, etc. ETG may be viewed also as a specific commitment. The nature of the SPE of ETG is then linked to the commitment robustness of the NE of the initial static game and to the value of commitment in this same game.

REFERENCES

- Albaek, S. (1990): "Stackelberg Leadership as a Natural Solution Under Cost Uncertainty," *The Journal of Industrial Economics*, 38(3), 335–347.
- Alexeev, M., and J. Leitzel (1996): "Rent Shrinking," *Southern Economic Journal*, 62, 620–626.
- Amegashie, J.A. (2001): "An All-pay Auction with a Pure-strategy Equilibrium," *Economics Letters*, 70(1), 79–82.
- Amir, M., R. Amir, and J. Jin (2000): "Sequencing R&D Decisions in a Two-period Duopoly with Spillovers," *Economic Theory*, 15(2), 297–317.

²² Moreover, we stress that the result of Morgan and Várdy (2007) relies on the uniqueness of the equilibrium in the sequential-move contest. A strategic uncertainty appears when the two Stackelberg equilibria are NE of the ETG. Such uncertainty may justify the use of risk dominance to select one of the two SPEs as in Van Damme and Hurkens (1997). This strategic uncertainty may restore the value of commitment even if observation is costly.

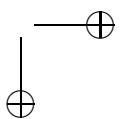
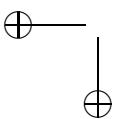
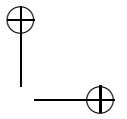
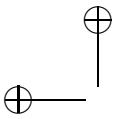


- Anbarci, N., S. Skaperdas, and C. Syropoulos (2002): "Comparing Bargaining Solutions in the Shadow of Conflict: How Norms Against Threats Can Have Real Effects," *Journal of Economic Theory*, 106(1), 1–16.
- Anderton, C.H., and J. Carter (2008): "Vulnerable Trade: The Dark Side of an Edgeworth Box," *Journal of Economic Behavior & Organization*, 68, 422–432.
- Anderton, C.H., R.A. Anderton, and J.R. Carter (1999): "Economic Activity in the Shadow of Conflict," *Economic Inquiry*, 37(1), 166–179.
- Appelbaum, E., and E. Katz (1987): "Seeking Rents by Setting Rents: The Political Economy of Rent-seeking," *Economic Journal*, 97(387), 685–699.
- Azam, J.-P. (1995): "How to Pay for the Peace? A Theoretical Framework with References to African Countries," *Public Choice*, 83(1), 173–184.
- Bade, S., G. Haeringer, and L. Renou (2009): "Bilateral Commitment," *Journal of Economic Theory*, 144(4), 1817–1831.
- Bagwell, K. (1995): "Commitment and Observability in Games," *Games and Economic Behavior*, 8, 271–280.
- Baik, K.H. (1994): "Effort Levels in Contests with Two Asymmetric Players," *Southern Economic Journal*, 61, 367–378.
- Baik, K.H., and I.-G. Kim (1997): "Delegation in Contests," *European Journal of Political Economy*, 13, 281–298.
- Baik, K.H., and J. Kim (2014): "Contest with Bilateral Delegation: Unobservable Contracts," *Journal of Institutional and Theoretical Economics*, 170(3), 387–405.
- Baik, K.H., and S. Lee (2007): "Collective Rent-seeking When Sharing Rules Are Private Information," *European Journal of Political Economy*, 23(3), 768–776.
- Baik, K.H., and D. Lee (2012): "Do Rent-seeking Groups Announce Their Sharing Rules?" *Economic Inquiry*, 50(2), 348–363.
- Baik, K.H., and J.H. Lee (2013): "Endogenous Timing in Contests with Delegation," *Economic Inquiry*, 51(4), 2044–2055.
- Baik, K.H., and J.F. Shogren (1992): "Strategic Behavior in Contests: Comment," *The American Economic Review*, 82(1), 359–362.
- Baik, K.H., and J.F. Shogren (1994): "Environmental Conflicts With Reimbursement for Citizen Suits," *Journal of Environmental Economics and Management (New York)*, 27, 1–20.
- Baik, K.H., T.L. Cherry, S. Kroll, and J.F. Shogren (1999): "Endogenous Timing in a Gaming Tournament," *Theory and Decision*, 47(1), 1–21.
- Barros, P.P., and L. Sørgard (2001): "Merger in an Advertising-intensive Industry," mimeo.
- Baumann, F., P. Denter, and T. Friehle (2013): "Hide or Show? Endogenous Observability of Private Precautions against Crime When Property Value is Private Information," *DICE Discussion Paper* 115, Düsseldorf.
- Baye, M.R., and H.C. Hoppe (2003): "The Strategic Equivalence of Rent-seeking, Innovation, and Patent-race Games," *Games and Economic Behavior*, 44(2), 217–226.
- Baye, M.R., D.J. Kovenock, and C.G. de Vries (2005): "Comparative Analysis of Litigation Systems: An Auction-theoretic Approach," *The Economic Journal*, 115, 583–601.
- Baye, M.R., D.J. Kovenock, and C.G. De Vries (2012): "Contests with Rank-order Spillovers," *Economic Theory*, 51(2), 315–350.
- Bernardo, A.E., E. Talley, and I. Welch (2000): "A Theory of Legal Presumptions," *Journal of Law, Economics, & Organization*, 16(1), 1–49.
- Beviá, C., and L.C. Corchón (2010): "Peace Agreements Without Commitment," *Games and Economic Behavior*, 68(2), 469–487.
- Bulow, J.I., J.D. Geanakoplos, and P.D. Klemperer (1985a): "Holding Idle Capacity to Deter Entry," *The Economic Journal*, 95, 178–182.
- Bulow, J.I., D. Geanakoplos, and P.D. Klemperer (1985b): "Multimarket Oligopoly: Strategic Substitutes and Complements," *The Journal of Political Economy*, 93(3), 488–511.
- Cai, H. (2003): "War or Peace," *Contributions to Economic Analysis & Policy*, 2(1), 1–28.
- Che, Y.-K., and I. Gale (2003): "Optimal Design of Research Contests," *The American Economic Review*, 93(3), 646–671.
- Chioveanu, I. (2008): "Advertising, Brand Loyalty and Pricing," *Games and Economic Behavior*, 64(1), 68–80.
- Chowdhury, S.M., and R.M. Sheremeta (2011): "Multiple Equilibria in Tullock Contests," *Economics Letters*, 112(2), 216–219.
- Chung, T.-Y. (1996): "Rent-seeking Contest When the Prize Increases with Aggregate Efforts," *Public Choice*, 87, 55–66.
- Clark, D.J., and C. Riis (2007): "Contingent Payments in Selection Contests," *Review of Economic Design*, 11(2), 125–137.
- Cohen, C., and A. Sela (2005): "Manipulations in Contests," *Economics Letters*, 86(1), 135–139.

- Cohen, C., and T. Shavit (2012): "Experimental Tests of Tullock's Contest With and Without Winner Refunds," *Research in Economics*, 66(3), 263–272.
- Cohen, C., T.R. Kaplan, and A. Sela (2008): "Optimal Rewards in Contests," *The RAND Journal of Economics*, 39(2), 434–451.
- Corchón, L.C. (2007): "The Theory of Contests: A Survey," *Review of Economic Design*, 11, 69–100.
- Dale, D.J. (2004): "Charitable Lottery Structure and Fund Raising: Theory and Evidence," *Experimental Economics*, 7(3), 217–234.
- Daughety, A.F., and J.F. Reinganum (1994): "Asymmetric Information Acquisition and Behavior in Role Choice Models: An Endogenously Generated Signaling Game," *International Economic Review*, 35(4), 795–819.
- Dechenaux, E., D. Kovenock, and R.M. Sheremeta (2014): "A Survey of Experimental Research on Contests, All-pay Auctions and Tournaments," *Experimental Economics*, 18(4), 609–669.
- De Frutos, M.-A., C. Ornaghi, and G. Siotis (2013): "Competition in the Pharmaceutical Industry: How Do Quality Differences Shape Advertising Strategies?" *Journal of Health Economics*, 32(1), 268–285.
- Deneckere, R.J., and D.J. Kovenock (1992): "Price Leadership," *The Review of Economic Studies*, 59, 143–162.
- Denter, P., and D. Sisak (2015): "The Fragility of Deterrence in Conflicts," *Journal of Theoretical Politics*, 27(1), 43–57.
- Dixit, A. (1987): "Strategic Behavior in Contests," *The American Economic Review*, 77(5), 891–898.
- Eaton, C.B. (2004): "The Elementary Economics of Social Dilemmas," *Canadian Journal of Economics*, 37(4), 805–829.
- Espinosa, M.P., and P. Mariel (2001): "A Model of Optimal Advertising Expenditures in a Dynamic Duopoly," *Atlantic Economic Journal*, 29(2), 135–161.
- Farmer, A., and P. Pecorino (1999): "Legal Expenditure as a Rent-seeking Game," *Public Choice*, 100, 271–288.
- Fehr, E., and K.M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114(3), 817–868.
- Fey, M. (2006): "Rent-seeking Contest with Incomplete Information," *Public Choice*, 129, 225–236.
- Fonseca, M.A. (2009): "An Experimental Investigation of Asymmetric Contests," *International Journal of Industrial Organization*, 27, 582–591.
- Friedman, J.W. (1983): *Oligopoly Theory*, Cambridge, UK: Cambridge University Press.
- Fu, Q. (2006): "Endogenous Timing of Contests with Asymmetric Information," *Public Choice*, 129, 1–23.
- Fu, Q., O. Gürler, and J. Münster (2013): "Communication and Commitment in Contests," *Journal of Economic Behavior & Organization*, 95, 1–19.
- Garfinkel, M.R., and S. Skaperdas (2000): "Conflict Without Misperceptions or Incomplete Information: How the Future Matters," *The Journal of Conflict Resolution*, 44(6), 793–807.
- Garfinkel, M.R., and S. Skaperdas (2007): "Economics of Conflict: An Overview," in T. Sandler, and K. Hartley (eds), *Handbook of Defense Economics in a Globalized World, Vol. II*, Amsterdam: North-Holland, pp. 649–709.
- Gershkov, A., J. Li, and P. Schweinzer (2009): "Efficient Tournaments Within Teams," *The RAND Journal of Economics*, 40(1), 103–119.
- Glazer, A., and R. Hassin (2000): "Sequential Rent-seeking," *Public Choice*, 102, 219–228.
- Gong, J., and R.P. McAfee (2000): "Pretrial Negotiation, Litigation, and Procedural Rules," *Economic Inquiry*, 38(2), 218–238.
- Gradstein, M. (1993): "Rent-seeking and the Provision of Public Goods," *The Economic Journal (New York)*, 103, 1236–1243.
- Gregor, M. (2012): "Contest for Power in Organizations," *Economics Letters*, 114(3), 280–283.
- Grossman, H.-I. (2001): "The Creation of Effective Property Rights," *The American Economic Review*, 91(2), 347–352.
- Grossman, H.I., and M. Kim (1995): "Swords or Plowshares? A Theory of the Security of Claims to Property," *The Journal of Political Economy*, 103(6), 1275–1288.
- Haan, M.A., and J.L. Moraga-González (2011): "Advertising for Attention in a Consumer Search Model," *The Economic Journal*, 121(552), 552–579.
- Hamilton, J.H., and S.M. Slutsky (1990): "Endogenous Timing in Oligopoly Games: Stackelberg or Cournot Equilibria," *Games and Economic Behavior*, 2, 29–46.
- Harsanyi, J.C., and R. Selten (1988): *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.
- Hausken, K. (2000): "Cooperation and Between-group Competition," *Journal of Economic Behavior & Organization*, 42, 417–425.
- Hirschleifer, J. (1988): "The Analytics of Continuing Conflict," *Synthese*, 76, 201–233.
- Hirschleifer, J. (1989): "Conflict and Rent-seeking Success Functions: Ratio vs. Difference Models of Relative Success," *Public Choice*, 63, 101–112.
- Hirschleifer, J. (1991a): "The Paradox of Power," *Economics and Politics*, 3(3), 177–200.

- Hershleifer, J. (1991b): "The Technology of Conflict as an Economic Activity," *The American Economic Review, Papers and Proceedings of the Hundred and Third Annual Meeting of the American Economic Association*, 81(2), 130–134.
- Hershleifer, J. (1995a): "Anarchy and its Breakdown," *The Journal of Political Economy*, 103(1), 26–52.
- Hershleifer, J. (1995b): "Theorizing about Conflict," in K. Hartley, and T. Sandler (eds), *Handbook of Defense Economics, Vol. I*, Amsterdam: Elsevier, pp. 165–189.
- Hirshleifer, J., and E. Osborne (2001): "Truth, Effort, and the Legal Battle," *Public Choice*, 108(1), 169–195.
- Hodler, R., and H. Yektaş (2012): "All-pay War," *Games and Economic Behavior*, 74(2), 526–540.
- Hoffmann, M. (2010): "Enforcement of Property Rights in a Barter Economy," *Social Choice and Welfare*, 34, 249–263.
- Hoffmann, M. (2015): "Multiple Rounds and Endogenous Timing in Endogenous Prize Contests with Asymmetric Valuation," mimeo.
- Hoffmann, M., and M. Kolmar (2013): "Distributional Preferences in Probabilistic and Share Contests," *CESifo Working Paper No. 4184*.
- Hoffmann, M., and G. Rota-Graziosi (2012): "Endogenous Timing in General Rent-seeking and Conflict Models," *Games and Economic Behavior*, 75, 168–184.
- Hoffmann, M., and G. Rota-Graziosi (2014): "Endogenous Timing in the Presence of Non-monotonicities," mimeo.
- Hotte, L. (2001): "Conflicts over Property Rights and Natural-resource Exploitation at the Frontier," *Journal of Development Economics*, 66(1), 1–21.
- Hurley, T.M. (1998): "Rent Dissipation and Efficiency in a Contest with Asymmetric Valuations," *Public Choice*, 94, 289–298.
- Kalai, A.T., E. Kalai, E. Lehrer, and D. Samet (2010): "A Commitment Folk Theorem," *Games and Economic Behavior*, 69(1), 127–137.
- Kaplan, T.R., I. Luski, A. Sela, and D. Wettstein (2002): "All-pay Auctions with Variable Rewards," *The Journal of Industrial Economics*, 50(4), 417–430.
- Kaplan, T.R., I. Luski, and D. Wettstein (2003): "Innovative Activity and Sunk Cost," *International Journal of Industrial Organization*, 21(8), 1111–1133.
- Kolmar, M. (2008): "Perfectly Secure Property Rights and Production Inefficiencies in Tullock Contests," *Southern Economic Journal*, 75(2), 441–456.
- Konrad, K.A. (2002): "Investment in the Absence of Property Rights: The Role of Incumbency Advantages," *European Economic Review*, 46(8), 1521–1537.
- Konrad, K.A. (2009): *Strategy and Dynamics in Contests*. New York: Oxford University Press.
- Konrad, K.A., and H. Schlesinger (1997): "Risk Aversion in Rent-seeking and Rent-augmenting Games," *The Economic Journal*, 107(445), 1671–1683.
- Lange, A., J.A. List, and M.K. Price (2007): "A Fundraising Mechanism Inspired by Historical Tontines: Theory and Experimental Evidence," *Journal of Public Economics*, 91(9), 1750–1782.
- Lee, S. (1995): "Endogenous Sharing Rules in Collective-group Rent-seeking," *Public Choice*, 85, 31–44.
- Leininger, W. (1993): "More Efficient Rent-seeking – A Münchhausen Solution," *Public Choice*, 75, 43–62.
- Leininger, W., and C.-L. Yang (1994): "Dynamic Rent-seeking," *Games and Economic Behavior*, 7, 406–427.
- Lim, B.I., and J.F. Shogren (2004): "Unilateral Delegation and Reimbursement Systems in an Environmental Conflict," *Applied Economics Letters*, 11(8), 489–493.
- Linster, B.G. (1993): "Stackelberg Rent-seeking," *Public Choice*, 77, 307–321.
- Mailath, G.J. (1993): "Endogenous Sequencing of Firm Decisions," *Journal of Economic Theory*, 59(1), 169–182.
- Matros, A. (2012): "Sad-loser Contests," *Journal of Mathematical Economics*, 48(3), 155–162.
- Matros, A., and D. Armanios (2009): "Tullock's Contest with Reimbursements," *Public Choice*, 141, 49–63.
- Moldovanu, B., and A. Sela (2001): "The Optimal Allocation of Prizes in Contests," *The American Economic Review*, 91(91), 542–558.
- Morath, F., and J. Münster (2013): "Information Acquisition in Conflicts," *Economic Theory*, 54(1), 99–129.
- Morgan, J. (2000): "Financing Public Goods by Means of Lotteries," *The Review of Economic Studies*, 67(4), 761–784.
- Morgan, J. (2003): "Sequential Contests," *Public Choice*, 116, 1–18.
- Morgan, J., and F. Várdy (2007): "The Value of Commitment in Contests and Tournaments When Observation is Costly," *Games and Economic Behavior*, 60(2), 326–338.
- Morgan, J., and F. Várdy (2013): "The Fragility of Commitment," *Management Science*, 59(6), 1344–1353.
- Münster, J. (2007): "Contests with Investment," *Managerial and Decision Economics*, 28(8), 849–862.
- Neary, H.M. (1997): "Equilibrium Structure in an Economic Model of Conflict," *Economic Inquiry*, 35, 480–494.
- Nitzan, S. (1991a): "Collective Rent Dissipation," *The Economic Journal*, 101(409), 1522–1534.
- Nitzan, S. (1991b): "Rent-seeking with Non-identical Sharing Rules," *Public Choice*, 71, 43–50.
- Nitzan, S. (1994): "Modelling Rent-seeking Contests," *European Journal of Political Economy*, 10, 41–60.
- Nitzan, S., and K. Ueda (2011): "Prize Sharing in Collective Contests," *European Economic Review*, 55, 678–687.

- Noh, S.J. (1999): "A General Equilibrium Model of Two Group Conflict with Endogenous Intra-group Sharing Rules," *Public Choice*, 98, 251–267.
- Park, S.-H. (2010): "Asymmetric Reimbursement System in an Environmental Conflict," *Applied Economics Letters*, 17(10/12), 1197–1199.
- Pérez-Castrillo, J.D., and T. Verdier (1992): "A General Analysis of Rent-seeking Games," *Public Choice*, 73(3), 335–350.
- Renou, L. (2009): "Commitment Games," *Games and Economic Behavior*, 66(1), 488–505.
- Ridlon, R. (2013): "Does Manager Effort Crowd-in or Crowd-out Workers' Efforts," mimeo.
- Robson, A.J. (1990): "Stackelberg and Marshall," *The American Economic Review*, 80(1), 69–82.
- Rosenthal, R.W. (1991): "A Note on Robustness of Equilibria with Respect to Commitment Opportunities," *Games and Economic Behavior*, 3, 237–243.
- Runkel, M. (2011): "Revenue Sharing, Competitive Balance and the Contest Success Function," *German Economic Review*, 12(3), 256–273.
- Saloner, G. (1987): "Cournot Duopoly with Two Production Periods," *Journal of Economic Theory*, 42(1), 183–187.
- Schelling, T.C. (1960): *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
- Schmalensee, R. (1976): "A Model of Promotional Competition in Oligopoly," *The Review of Economic Studies*, 43, 493–507.
- Shaffer, S. (2006): "Contests with Interdependent Preferences," *Applied Economics Letters*, 13, 877–880.
- Shapiro, C. (1989): "Theories of Oligopoly Behavior," in R. Schmalensee, and R.D. Willi (eds), *Handbook of Industrial Organization, Vol. I*, Amsterdam: Elsevier, pp. 329–414.
- Shogren, J.F., and K.H. Baik (1992): "Favorites and Underdogs: Strategic Behavior in an Experimental Contest," *Public Choice*, 74(2), 191–205.
- Shogren, J.F., and T.M. Hurley (1997): "Tournament Incentives in Environmental Policy," in A.K. Dragun, and K.M. Jakobsson (eds), *Sustainability and Global Environmental Policy: New Perspectives*, Cheltenham, UK and Lyme, NH, USA, pp. 213–231.
- Skaperdas, S. (1992): "Cooperation, Conflict, and Power in the Absence of Property Rights," *The American Economic Review*, 82(4), 720–739.
- Skaperdas, S., and C. Syropoulos (2002): "Insecure Property and the Efficiency of Exchange," *The Economic Journal*, 112, 133–146.
- Smith, A.C., D. Houser, P.T. Leeson, and R. Ostadhossein (2014): "The Costs of Conflict," *Journal of Economic Behavior & Organization*, 97, 61–71.
- Spier, K.E. (2007): "Litigation," in A.M. Polinsky, and S. Shavell (eds), *Handbook of Law and Economics*, Amsterdam: North-Holland, pp. 259–342.
- Sun, G.-Z., and Y.-K. Ng (1999): "The Effect of Number and Size of Interest Groups on Social Rent Dissipation," *Public Choice*, 101, 251–265.
- Szymanski, S. (2003): "The Economic Design of Sporting Contests," *Journal of Economic Literature*, 41(4), 1137–1187.
- Tullock, G. (1980): "Efficient Rent-seeking," in J. Buchanan, R. Tollison, and G. Tullock (eds), *Towards a Theory of the Rent-seeking Society*, College Station, TX: A&M University Press, pp. 97–112.
- Tullock, G. (1985): "Efficient Rents 3 – Back to the Bog," *Public Choice*, 46(3), 259–263.
- Van Damme, E., and S. Hurkens (1996): "Commitment Robust Equilibria and Endogenous Timing," *Games and Economic Behavior*, 15, 290–311.
- Van Damme, E., and S. Hurkens (1997): "Games with Imperfectly Observable Commitment," *Games and Economic Behavior*, 21, 282–308.
- Vives, X. (2001): *Oligopoly Pricing – Old Ideas and New Tools*. Cambridge, MA: MIT Press.
- Vogt, C., J. Weimann, and C.-L. Yang (2002): "Efficient Rent-seeking in Experiment," *Public Choice*, 110, 67–78.
- von Stackelberg, H. (1934): *Marktform und Gleichgewicht*. Vienna: Julius Springer.
- Wärneryd, K. (2000): "In Defense of Lawyers: Moral Hazard as an Aid to Cooperation," *Games and Economic Behavior*, 33(1), 145–158.
- Wärneryd, K. (2003): "Information in Conflicts," *Journal of Economic Theory*, 110(1), 121–136.
- Weimann, J., C.-L. Yang, and C. Vogt (2000): "An Experiment on Sequential Rent-seeking," *Journal of Economic Behavior & Organization*, 41(4), 405–426.
- Yildirim, H. (2005): "Contests with Multiple Rounds," *Games and Economic Behavior*, 51, 213–227.
- Zhou, H. (2006): "R&D Tournaments with Spillovers," *Atlantic Economic Journal*, 34(3), 327–339.

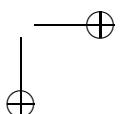
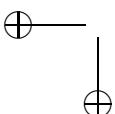
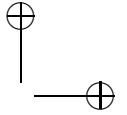
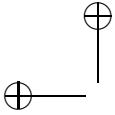


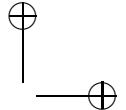
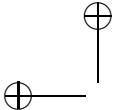


PART III

SPECIAL TOPICS







8. Firm pricing with consumer search

*Simon P. Anderson and Régis Renault**

1 INTRODUCTION

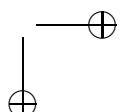
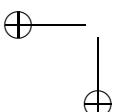
Consumer search in the modern economy seems more prevalent than ever before, with the advent of Internet shopping opportunities. The lower per search cost enabled through the Internet, concurrent with low transactions costs for shipping final purchases, has enabled consumers to access a huge variety of options. Heretofore the effective choice set was limited by the local market, and even shopping in that market was constrained by quite high costs of becoming informed as to what products were offered at what prices. Search costs have always been important, but, since few people actually actively searched much, it was perhaps not apparent that search costs have a major constraining role in stifling economic activity. Ironically, because search costs were less visible, economists often ignored them and addressed seemingly more pressing questions about pricing, product variety, and market power, while treating the set of available products as effectively given and known.

Now that search is visibly at the fore of the modern economy, economists are developing (or redeveloping) models of market interaction in which search costs play a central role. Search costs are a friction in markets, and their reduction facilitates transactions that are likely to involve both better matching (of consumer to product) and lower prices (because there is more effective competition). One key reason why there is now so much more accessible choice is that in earlier times high search costs effectively deterred access to available options: search costs silently curtailed the set of goods and services brought to market. Nonetheless, significant search frictions remain.

To set the stage from the perspective of economic theory of oligopoly pricing, we start by describing how costly search impacts economic outcomes in the simplest scenario, which we can then embellish to address more subtle features. The first result is both striking and quite extreme.

Assume that firms set prices for a homogeneous good, and produce at a common constant marginal cost. Absent search costs, this market interaction gives us the Bertrand paradox (itself a striking and extreme result): *price is at marginal cost* as long as at least two firms compete. Now introduce a small search cost, s , for the consumer to find out the price set by a firm, but assume she observes a first price quote for free. Suppose that consumers search sequentially among options, and close the model with the rational expectations stipulation that they are correct in equilibrium about the price(s) they expect. To make it simple, suppose that the consumers all have the same unit demand curve, with maximal valuation r for the product. Then, the outcome – the so-called “Diamond paradox” after its first proponent (Diamond, 1971) – has all firms pricing at the *monopoly price*, r . Roughly, one can think of prices being

* We thank Mark Armstrong, José Luis Moraga-González, Andrew Rhodes, Vaiva Petrikaitė and Jidong Zhou for invaluable comments. We also thank Emily Cook for admirable research assistance. The first author thanks the NSF for financial support; the second author thanks Labex MME-DII for financial support.



“ratcheted up” (by the amount of the search cost) for any putative lowest price below the monopoly level, r . (In the text, we develop and extend the arguments and contexts that give rise to the paradox.) Thus even a small search cost radically tips the Bertrand paradox to the opposite extreme.¹ There is a second leg to the paradox: the context is a search model, and yet no one actually searches again in equilibrium. This is logical because prices are rationally expected to be the same for all firms. And it is because there is no search that the monopoly price sustains.²

To be sure, economists were thinking about search before Diamond’s market equilibrium solution. Stigler (1961) was intrigued by the price dispersion (and the implicit repealing of the law of one price) that he saw in markets from anthracite coal to Chevrolets (see also Lach, 2002, for a study documenting price dispersion in Israeli supermarkets for instant coffee, frozen chickens, flour, and refrigerators; Ellison and Fisher Ellison, 2014, for online books, etc.). This led Stigler to begin formulating consumer search theory (we develop this theme below using the formal model of Burdett and Judd, 1983 as our jumping-off point for the formal analysis with equilibrium pricing). The theory of consumer search – with the consumer facing a set of options with exogenous utility distributions – was developed quite briskly. It culminated in the beautiful characterization of the consumer problem in Weitzman (1979) whereby consumers search remaining options in order of their reservation utility scores until they reach a stopping point at which remaining options have scores below their current holding. This rule, and some of the work that led up to it, is discussed in more detail below.

Somewhat ironically though, even though Stigler felt that consumer search frictions should be at the heart of observed price dispersion, the Diamond result suggested otherwise – at least under the assumption that consumers search sequentially across options (we discuss simultaneous search with reference to Stigler in the next section). There are, of course other reasons for price dispersion, such as heterogeneity in consumer tastes across products, and cost differences across firms, but search costs are not one of them – at least in the Diamond set-up. This observation leads us to consider what could generate price dispersion and equilibrium search when search is indeed sequential (contrast Stigler on this point), and products are homogeneous to consumers (i.e., absent product differentiation). As we argue below, pure production cost heterogeneity is insufficient (Reinganum, 1979): while this generates equilibrium price dispersion (with a pooling atom at the price that just deters further search) it generates no search. Demand heterogeneity alone is also unlikely to give dispersion, and in many instances pure search cost heterogeneity alone does not either (Rob, 1985, Stahl, 1996). However, there is a notable exception, which arises when some consumers have zero search costs. Indeed, Stahl (1989) has provided an enduring work horse model of pricing under consumer search that generates price dispersion. In this model, products are homogeneous but a fraction of consumers have strictly positive search costs while the rest are assumed to

¹ Some classic ways of softening the Bertrand paradox (see e.g., Tirole, 1988) include introducing product differentiation, repeated interaction, and decreasing returns to scale. We will discuss below how product differentiation relaxes the Diamond paradox. The relevant repeated interaction in the search context would be long-term relationships between buyers and sellers with repeat purchases, where prices would be kept low in order to ensure ongoing business in the future. Allowing for decreasing returns (increasing marginal costs) with consumers still searching randomly would imply that “Diamond” equilibrium prices decrease with the number of firms in the market, as the “monopoly” market is divided into more pieces. Conversely, locally increasing returns (not so severe as to violate second-order conditions) cause prices to *rise* with entry.

² This argument, as we shall see below, is why product differentiation breaks the paradox: people search for better product matches, and thus firms are brought into competition with each other.

observe all prices (and they can be construed as a mass of consumers with zero search costs).³ The equilibrium is in mixed strategies, and constructing it (via an expository device suggested by Janssen, Moraga-González, and Wildenbeest, 2005) builds off the ingredient models of Varian's (1980) model of sales coupled to the constraint that the consumer reservation price for the high-cost types just deters them from searching (engaging one striking consequence of Weitzman's search rule – that if a myopic search is not optimal, then it is optimal not to search at all).⁴

Allowing for differentiated products is an appealing way to break the Diamond paradox. Wolinsky (1986) first explored this avenue, and Anderson and Renault (1999) develop the economics of the approach further. The idea is that search is not just for price but also for product suitability with a consumer's tastes, and different products suit different consumers better ("horizontal differentiation"). In this case a consumer may well consider searching even if she encounters the same price she expects elsewhere (or even if she encounters a lower one), because she may seek a better match to her desired product specification. The situation is most tractably modeled as a differentiated product discrete choice model with independent and identically distributed (i.i.d) valuations (idiosyncratic taste matches) across consumers. Under symmetric demands, there is a symmetric equilibrium price, so this approach does not per se deliver price dispersion.⁵ It does nonetheless deliver consumer search in equilibrium: different consumers search different numbers of options, even if their search costs are the same (because some are luckier in finding satisfactory matches early). Moreover, higher search costs induce less search, which in turn by increasing friction entails less effective competition and hence higher equilibrium prices. Another result ties together nicely the standard product differentiation literature with the Diamond paradox reasoning. In the standard product differentiation setting (with no search frictions), a higher variance of consumer tastes for products (i.e., more heterogeneity of products) leads to higher equilibrium prices as individuals have more pronounced idiosyncratic tastes and so individual product demands are more inelastic. The presence of search costs can cause equilibrium prices to initially decrease as the degree of product differentiation rises. The key here is that more product differentiation leads to more search, which leads to more competition as more firms are effectively brought into the contest for a consumer. At first (i.e., for low product heterogeneity), this causes prices to fall down from the high Diamond levels (where there is no search at all). For higher heterogeneity consumers are searching sufficiently that they know about a significant fraction of firms. Then this is like the model without search costs, and more differentiation begets higher prices through greater product loyalty.

In many contexts – from geographical markets through online shopping – search is both sequential and systematically ordered. Ordering gives prominence to firms searched early on, and can quite drastically influence equilibrium pricing. Because earlier firms get more traffic they are usually more profitable, so firms are willing to expend resources (bid for prime locations or high positions in search engine auctions) for better positions.

³ Because the high search cost types never search again in equilibrium, one might argue that there is no real search in this model.

⁴ Varian's (1980) model of sales is described in the next section. Weitzman's (1979) key results are in subsection 4.4.

⁵ Different prices can be generated by allowing firms to have different production costs or vertical qualities, at the cost of a significantly higher consequent complication of the analysis. Another important channel for generating price dispersion is to have ordered consumer search, as discussed at length below.

Equilibrium orders depend on differential pay-offs at different positions. When firms are intrinsically heterogeneous and asymmetric (e.g., when they command different distributions of consumer valuations), the equilibrium order is driven by profitability differences. However, profit differences do not necessarily reflect consumer welfare differences. Thus the order of presentation of options resulting from an auction might be unattractive to buyers who might consequently be put off from participating on a platform (Anderson and Renault, 2016).

We devote the last subsection within each section to ordered search, so as to discuss its impact in the various contexts of product heterogeneity, seller heterogeneity, and buyer heterogeneity. Contemporaneous work by Armstrong (2016) both surveys and extends the pricing theory in this important direction for the case of product heterogeneity, and we refer the reader to Armstrong's excellent paper for further elaboration.⁶

One organizing theme we stress throughout is the impact of reducing search costs, as predicted by various approaches to modeling pricing with search costs. Our motivation, as per the start of this introduction, is the Internet experience and its impact on markets. We also take a broader perspective by engaging the endogeneity of the consumer (and firm) participation decision, recognizing the two-sided nature of many search markets. The thickness of the market may be greater with lower search costs, and must be accounted for to fully appreciate the benefits of search cost reductions.

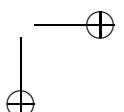
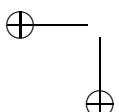
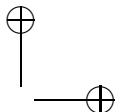
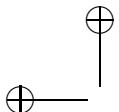
2 PRICE DISPERSION

The central argument in Stigler (1961) is that price dispersion may persist because consumers are not aware of all prices and therefore cannot properly arbitrage price differences. Here we present various settings where costly search by buyers generates price dispersion resulting from the price mixed strategies used by firms. We focus more specifically on settings where such price dispersion arises in spite of a homogeneous population of sellers and a homogeneous population of buyers. Baye, Morgan, and Scholten (2006) provide a fine survey of models of price dispersion with a homogeneous product, and emphasize the role of a clearinghouse that publishes prices. To fix ideas about how mixed strategies arise when consumers are imperfectly informed about prices we start, however, with a simple setting that allows for some heterogeneity on the demand side.

2.1 Imperfect Buyer Information Yields Price Dispersion

Consider a market comprised of $n > 1$ firms with zero production costs selling a homogeneous product to a continuum of consumers with unit demand. A consumer's valuation is denoted $r > 0$. The total measure of consumers is $m > 0$, out of which there is a measure $\sigma \in [0, m]$ of shoppers who observe all prices at no cost. Shoppers therefore always buy at the lowest price in the market. The remaining consumers observe only one price for free and face a prohibitively high search cost if they wish to observe an additional price quote. These captive consumers therefore always buy from the firm whose price quote they get for free, provided that its price does not exceed r . They are equally shared among the firms, and $\gamma = \frac{m-\sigma}{n}$

⁶ One particularly stimulating contribution of Armstrong's paper is the reformulation of sequential search as a static discrete choice problem without search frictions.



denotes the measure of captive consumers per firm. Hence a firm is guaranteed a profit of γr , which would result from charging the monopoly price r and selling exclusively to captive consumers.

A standard Bertrand undercutting argument shows that there cannot be a pure strategy equilibrium where shoppers pay a price that strictly exceeds marginal cost: a firm sharing the shoppers' demand with some other firm(s) at a price above marginal cost could always profitably slightly undercut that price to capture the entire shoppers' demand. However, pricing at marginal cost cannot be an equilibrium in the current setting because a firm charging that price would earn zero profit, whereas it can always guarantee a strictly positive profit by giving up selling to shoppers altogether and extracting the entire surplus from its captive consumers.

The literature, following Varian (1980), has focused on characterizing the unique symmetric mixed strategy equilibrium of this game. As will be seen below, the characterization of this equilibrium turns out to be quite relevant to the analysis of more elaborate search frameworks. The undercutting argument mentioned above can be used again to show that the price distribution can have no atom. Furthermore, a firm charging the highest price in the support of the price distribution will sell to shoppers with probability zero, therefore it is optimal for that firm to charge the captive consumers' valuation, r . The Bertrand intuition that all profits from sales to shoppers are competed away carries over here, and a firm's equilibrium profit is equal to its fall-back profit from selling to captive consumers alone, γr . However, a firm's price cannot fall below the level at which selling to all shoppers plus its captive consumers equals the fall-back profit. Thus, the minimum equilibrium price, \underline{p} , solves $(\gamma + \sigma)\underline{p} = \gamma r$, so $\underline{p} = \frac{\gamma}{\gamma + \sigma}r$. The equilibrium price distribution is obtained by ensuring that a firm maximizes its profit at all prices within the support and hence earns the same profit at all those prices. Appendix A shows that the equilibrium price distribution function is given by

$$F(p) = 1 - \left[\frac{\gamma}{\sigma} \left(\frac{r}{p} - 1 \right) \right]^{\frac{1}{n-1}}. \quad (8.1)$$

As should be expected, prices are stochastically higher (i.e., $F(p)$ is lower over the entire support, reflecting first-order stochastic dominance) if either the consumer valuation, r , or the relative share of captive consumers (reflected in $\frac{\gamma}{\sigma}$) are higher.

The comparative statics with respect to the number of firms is, however, less intuitive, and depend upon whether the new firm brings its new captive consumers with it when entering. If it does so, the fraction of shoppers in the consumer population falls. This is the situation described in Rosenthal (1980). Alternatively, the fraction of shoppers can be held fixed, and an entrant could just garner its share of the captive population. This is the situation germane to the search context we elaborate upon later. In both cases, there are two relevant expected prices to track. The first is the expected price per firm generated from $F(p)$: this is the price paid by captive consumers. The second is the expected minimum price, which is the price paid by the shoppers. Notice that more firms means more draws from the price distribution, which per se drives a lower price. However, for both cases (whether total shoppers or total captives are constant) $F(p)$ rises, so that what happens to the expected shopper price depends on which effect dominates.

First consider the situation analyzed by Rosenthal (1980), which is equivalent to letting the population of captive consumers increase so as to keep γ unchanged as more firms enter the

market.⁷ Then prices are first-order stochastically higher with more firms (to see this, note that for prices in the interior of the support $[\underline{p}, r]$, the bracketed term is in $[0, 1]$ so $F(p)$ falls). This reflects the logic of mixed strategies that requires that, as the number of competitors increases, a firm must find it as profitable to drop its price below r to attract shoppers as to charge r to its captive consumers. Profit from the latter choice is independent of the number of firms, so there is more reason to focus on the captive (or “loyal”) ones because there is less chance to be the low-price seller when there are more rivals. Hence the expected price paid by captive consumers rises. Rosenthal (1980) shows that this effect dominates the effect of more draws, so the price to shoppers goes up too.⁸ Baye et al. (2006) note that this might be expected because the shoppers are getting smaller as a fraction of the market.

Now suppose that the shopper population remains constant so that γ is decreasing in n . Janssen and Moraga-González (2004) show that the expected captive price rises (although the distributions are not first-order stochastically ordered). However, as shown by Morgan, Orzen, and Sefton (2006), the expected shopper price falls with entry,⁹ as the shoppers get relatively more important to individual firms, and the effect of more price draws dominates. Thus the prices move in different directions. As we shall see in subsection 5.2, when the reservation price is endogenously derived from optimal search behavior, the expected shopper price is independent of n .

The price dispersion that arises because of the captive consumers’ limited information should naturally feed into the incentives of these consumers to search on. This can only be captured in a setting where search costs are low enough in order for non-shoppers to have a genuine choice of whether to search or not. This possibility is explored in subsection 5.2 below. In the remainder of this section, we discuss how price dispersion may arise in settings where all consumers are identical and the heterogeneity in consumer information arises endogenously.

2.2 Simultaneous Search and the Coordination Problem

A first view of how consumers make their search decisions, which is consistent with Stigler’s original analysis (Stigler, 1961), is to assume that consumers commit to getting several price quotes before deciding where to buy. For instance, each consumer could choose to request price quotes that would be sent back in due time. To analyze this situation, we modify the setting above by assuming that all consumers get a first price quote for free, and then incur a cost $s > 0$ for each additional price quote requested. The set of firms’ prices observed by a consumer is selected randomly.

First remark that there is no equilibrium¹⁰ where all consumers request at least one additional price quote. Indeed, if all consumers observed at least two prices, then all firms would price at marginal cost: each firm competes à la Bertrand for each consumer with all other firms whose price is observed by that consumer. Expecting this though, no consumer

⁷ Rosenthal (1980) allows for general demand.

⁸ Rosenthal (1980) also notes the supports of the equilibrium prices do not change with n .

⁹ A quick confirmation of this result follows from the analysis in Appendix D. There we have that the expected shopper price is constant under entry in a situation where the reservation price rises with entry. But when the reservation price is exogenous, as in the Varian (1980) model, the shopper price therefore falls under entry.

¹⁰ In this section we do not dwell on the equilibrium concept although we should be noted that we always assume that consumers have passive beliefs after observing a price deviation. Somewhat more formal discussions can be found in subsections 3.2 and 4.2.

would be willing to incur the cost of observing a second price. Hence, in equilibrium, there must be some consumers who see only one price. Importantly, no matter how small the search cost, and akin to the Diamond paradox noted in the introduction, there is always an equilibrium where all consumers see only one price and firms charge the monopoly price. Charging the monopoly price is clearly profit maximizing if consumers observe only one price, and it is optimal for consumers not to incur any search cost if they expect the same price at all firms.¹¹

Burdett and Judd (1983) explore the possibility of equilibria where consumers search beyond one firm and prices are dispersed. They consider symmetric equilibria where all firms choose the same probability distribution over prices. They show that there is no equilibrium where consumers observe more than two prices. Relevant equilibria are then those where consumers mix between observing only one price, where v_1 denotes the probability of such a choice, and requesting a second price quote, which they do with complementary probability $v_2 = 1 - v_1$. The firms' pricing behavior depends on these probabilities and must be such that consumers are indifferent between the two choices in equilibrium.

To characterize a firm's profit-maximizing behavior note first that the only consumers that are relevant to a firm's pricing decision are those who observe that firm's price. There are $\frac{2mv_2}{n}$ such consumers who also obtain a price quote from some competing firm and $\frac{mv_1}{n}$ consumers who only see the firm's price and are therefore in an analogous situation to captive consumers in the preceding subsection. Furthermore, consumers observing two prices are in the same position as "shoppers" in our analysis above when there are only two firms in the market. Now, for each consumer, the firm competes with at most one other competitor, so that it is readily seen that its profit expression given its expectation about that other firm's behavior is exactly that of subsection 2.1 with two firms, $n = 2$, with $\gamma = \frac{mv_1}{n}$ and $\sigma = \frac{2mv_2}{n}$. As shown in Appendix A the equilibrium price distribution function is therefore

$$F(p) = 1 - \left[\frac{v_1}{2v_2} \left(\frac{r}{p} - 1 \right) \right]. \quad (8.2)$$

The value of v_1 , and hence that of v_2 , are obtained from the requirement that a consumer should be indifferent between observing only one price, thus paying the expected price Ep , and incurring search cost s to see a second price with a corresponding expected payment Ep_{min} , where p_{min} is the minimum of two prices. We therefore need v_1 to solve

$$Ep - Ep_{min} = s. \quad (8.3)$$

Burdett and Judd (1983) show that the left-hand side, which measures the benefit from search, is quasiconcave in v_1 and zero at both ends of the $[0, 1]$ interval.¹² Intuitively, if v_1 is close to zero so that nearly all consumers observe two prices, prices are close to marginal cost with a high probability and the expected minimum of two prices does not differ much from the expected price. Towards the other extreme, if v_1 is close to 1 so that almost all consumers are

¹¹ This would not be an equilibrium if the first price quote was costly because consumers would not be willing to incur the cost, expecting to be held up at the monopoly price. However, if consumer demand is price sensitive as in Burdett and Judd (1983), then this equilibrium survives as long as the search cost is not too large.

¹² Armstrong, Vickers, and Zhou (2009b) show that it is concave. These authors analyze a generalized version of the Burdett and Judd (1983) model.

captive, prices are very likely to be close to the monopoly value r and so there is not much benefit from seeking a second price quote. Hence, if the search cost s is not too large, there are two values of v_1 and therefore two equilibria. One has low search intensity (v_1 large) and high prices (although this one is unstable: see the discussion in Fershtman and Fishman, 1992); the other one has high search intensity (v_1 low) and low prices.

An interesting property of these equilibria is that the number of sellers has no impact on the competitiveness of the market. It is quite intuitive that, because each firm expects to be competing with at most one other firm for each consumer, it behaves as if it were in a duopoly market, independent of the total number of competitors. Anderson et al. (1992) find an analogous result for a simultaneous search model with horizontal product differentiation, where search uncovers the consumer's match with the firm's product.¹³ They find that, as the search cost increases, consumers sample fewer firms, and the equilibrium price is merely the full information oligopoly price for a market with a number of competitors equal to the sample size selected by consumers.¹⁴

The setting in Burdett and Judd (1983) is coherent with that of Stigler (1961). Although Stigler argued that price dispersion existed because consumers could not compare all prices, he did not explicitly analyze how the price distribution endogenously results from the consumer's search behavior. Performing the full equilibrium analysis shows that, although price dispersion and search may emerge as an equilibrium outcome, there may also be a one-price equilibrium with no search beyond a first (monopoly) price quote.¹⁵ In order for the outcome conjectured by Stigler to arise, it is necessary that firm and consumer anticipations are coordinated: consumers search because they expect price dispersion and firms use mixed price strategies because they expect some consumers to observe more than one price.

Such a coordination problem raises the question of the role of intermediaries. Do they facilitate coordination? Do they improve social welfare? The work by Baye and Morgan (2001) provides some answers in a framework that shares many similarities with the one we have just analyzed.¹⁶ In their setting, consumers can join an Internet platform by incurring a subscription fee, in which case they observe all prices posted by participating firms. If they stay out, consumers can buy from a local firm. The two differences with the Burdett and Judd (1983) are that: (i) the number of prices observed by the consumer if she incurs the search cost is not chosen by the consumer but is determined by the firms' participation decisions; (ii) firms incur an advertising cost (set by the platform) to post a price on the platform, so this is a model of costly price advertising. Baye and Morgan characterize an equilibrium where all consumers choose to join the platform and firms play a mixed strategy in which they either join the platform and pick a price according to a distribution with support up to the monopoly price, or else they do not advertise and charge the monopoly price. Obviously, this two-sided market setting does not eliminate the coordination problem: as is standard, agents on each side participate only if they expect sufficient participation by agents on the other side (this is the classic "failure to launch" problem: more generally, coordination issues are addressed in

¹³ The next section discusses such search environments in detail for sequential search.

¹⁴ Contrary to results in Burdett and Judd (1983), the number of firms sampled can increase beyond three and equals the total number of firms in the market if the search cost is low enough. This is because the incentive to search is determined by the exogenous match value distribution rather than by an endogenous price distribution.

¹⁵ Armstrong et al. (2009b) show that a price cap in a Burdett-Judd type of market reduces the incentives to search (the left-hand side of (8.3) above is reduced) and this reduces the fraction of consumers who get two quotes. In turn, this actually increases the average price paid in the market.

¹⁶ See also Baye et al. (2006), which embeds the broader context and provides useful empirical evidence.

the voluminous literature on the “chicken and egg” problem, following Caillaud and Jullien, 2003). Furthermore, Baye and Morgan show that the intermediary can worsen social welfare if individual consumer demand is not sufficiently price elastic. In a specification where demand elasticity is small enough, it is never socially optimal that such a platform is created if it involves any set-up cost, because the social welfare benefit from the resulting decrease in price is too small. Yet, an intermediary can profitably create such a platform through which it extracts some of the producer and consumer surplus through the subscription fee and the advertising fee.

Whereas the results presented so far may be viewed as a partial validation of Stigler’s conjecture, they assume that search is “simultaneous,” which may hold when there are response lags as with getting builders’ quotes for one’s deck extension, but does not seem consistent with many modern actual search environments (including Internet search). As we now show, viewing search as sequential can lead to a drastic rejection of Stigler’s analysis.

2.3 Sequential Search and the Diamond Paradox

Assume now that, in the setting outlined in the previous subsection, consumers may choose whether or not to incur the search cost s to obtain a new price quote after observing a firm’s price (up to the point where they have observed all firms’ prices). First, one can see that the argument provided above (for simultaneous search) also shows that there always exists a monopoly pricing equilibrium at which there is no search when search is sequential. The drastic difference between the two environments arises when we look for the possibility of other equilibria.

For the simple unit demand specification, there is no scope for pricing above the consumer valuation r . Consider now a situation where firms charge prices strictly below r . If this were an equilibrium, consumers know that the lowest such price, denoted here \underline{p} , is the lowest price they can hope for and will therefore always buy if they encounter that price. Now, a firm charging \underline{p} could increase its price slightly by some amount less than $\min\{s, r - \underline{p}\}$: no consumer would choose to search on, anticipating that the best she could find is another firm charging \underline{p} . Hence there cannot be an equilibrium where \underline{p} is charged in equilibrium.¹⁷ The argument can readily be generalized to price elastic demand as long as monopoly profit is single peaked.¹⁸

This result has come to be called the Diamond paradox after Diamond (1971). It is puzzling on two counts. First, monopoly pricing prevails even if the search cost is very small or the number of competitors in the market is very large. Second, consumers never search in equilibrium. Both predictions are at odds with casual observation. The rest of this chapter discusses how the introduction of some heterogeneity among agents on either side of the market might resolve or temper the puzzle.

As we noted in our discussion of simultaneous search, the monopoly pricing equilibrium cannot exist if individual demand is completely inelastic and the first price quote is costly. As noted by Stiglitz (1979), this also holds true for sequential search and (as we explain below) even if consumers have heterogeneous valuations. Thus search costs, no matter how

¹⁷ For mixed strategies, think of \underline{p} as the minimum of the support of a firm’s mixed strategy; the argument then applies for a firm charging a price in the neighborhood of \underline{p} .

¹⁸ See Renault (2016) for a sketch of the proof.

small, lead to market unraveling. However, an active market can exist if individual demand is sufficiently price elastic so that consumer surplus at the monopoly price is strictly positive.

2.4 Price Advertising

In his seminal article, Butters (1977) argued that price advertising is an important channel through which the Diamond paradox can be overcome. He also makes the point that the cost of advertising combined with buyer imperfect information generates price dispersion.¹⁹ To illustrate the point, consider a simple setting where each firm may choose to inform *all* consumers about its price at some fixed cost A , and the advertising decision and price are chosen simultaneously by firms. If a consumer is informed about a firm's price through advertising, she may buy from it at no additional cost. Otherwise, we retain the setting above except that there are only two firms.²⁰

First, if both firms advertised with probability 1, then the outcome would be Bertrand competition with marginal cost pricing, which would yield a profit of $-A$, whereas a firm can ensure at least zero profit if it does not advertise. Second, if neither firm advertises, the outcome is the Diamond outcome with both firms charging r .²¹ However, if $A < \frac{mr}{2}$, each firm could profitably deviate by advertising and slightly undercutting r to capture the entire market. Hence, in a symmetric equilibrium firms necessarily mix between advertising and not advertising. We now describe a symmetric equilibrium.

We seek an equilibrium where consumers expect a firm that does not advertise to charge the monopoly price r . A firm must be indifferent between not advertising and charging a price of r , and advertising some price p in the support of the equilibrium pricing strategy conditional on advertising. Because no advertised price will exceed the valuation r , a firm that does not advertise sells only if the competitor does not advertise either (with the tie-breaking rule that, when indifferent, a consumer picks the advertised price). We show in Appendix A3 that equilibrium profit is A and the probability of advertising a price at or below p is given by

$$F(p) = 1 - \frac{2A}{mp}, \quad (8.4)$$

for $p \in \left[\frac{2A}{m}, r\right]$. Hence firms do not advertise (charging price r) with probability $\frac{2A}{mr}$. Note that a firm that does not advertise could not earn more by charging a price below r , so this characterizes a Nash equilibrium. This equilibrium has the interesting property that firm profit does not depend on search costs but rather is entirely determined by the advertising cost. This result does not hold in the more elaborate model of Robert and Stahl (1993), where each consumer faces a non-trivial search problem because she is not reached by an ad from all firms that advertise: she must therefore update her beliefs about the pricing behavior of those firms from which she has received no ad. In that setting, a lower search reduces prices and

¹⁹ See Renault (2016) for a detailed discussion of the main contributions in the literature following Butters (1977).

²⁰ Janssen and Non (2008) consider a similar advertising technology in a more elaborate setting.

²¹ Equivalently, the model applies in a non-search setting when r is an exogenous common posted price set by firms and is therefore the default price if consumers get no ad. Then advertising a price corresponds to offering a discount. As we show below, there is a positive probability of not advertising, and therefore a positive probability of transactions at the “regular” price. Anderson, Baik, and Larson (2016) analyze a similar model with heterogeneous consumer valuations, and render endogenous the posted prices too.

profits. However, prices do not tend to marginal cost as the search cost goes to zero. By contrast, driving advertising costs to zero does lead to marginal cost pricing, which is also the case in our simple setting.

Finally and importantly, although this equilibrium exhibits some price dispersion, consumers never search in equilibrium. This is also the case in the analysis of Robert and Stahl (1993) despite the influence of the search cost on prices. As we argue below, it is actually quite challenging to avoid such an outcome in a setting with homogeneous products unless we allow for heterogeneity on both sides of the market. Such two-sided heterogeneity arises quite naturally in settings with horizontally differentiated products, to which we now turn.

3 MATCHING PRODUCTS TO CONSUMERS

By considering consumer search in markets for horizontally differentiated products, Wolinsky (1984, 1986) initiated a very novel and fruitful approach. Up to now, and in line with the search literature up to the 1980s, our central question has been the emergence of price dispersion, which is needed to create some motive for consumers to engage in costly search in a market for a homogeneous product. Horizontal product differentiation introduces an alternative motive to search, which results from the consumer's desire to find a satisfactory taste match with the purchased product. We start with a first look at the optimal search problem in the simple case where all the alternatives are *a priori* equally attractive to search.

3.1 Optimal Search Behavior

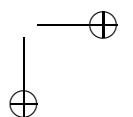
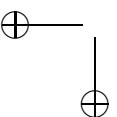
A consumer chooses one among n alternatives, where alternative i is expected to yield utility u_i , where u_i is a random variable i.i.d. across alternatives: it could be, for instance, that all products are sold at the same price and match utilities are i.i.d. or, as in the previous section, a homogeneous product is sold by firms playing a symmetric mixed strategy in prices. Let G denote the distribution function of u_i . The consumer may uncover her true utility with alternative i by incurring a search cost $s > 0$. At any time, she may enjoy any of the alternatives for which she has learned her utility, with no additional cost: this is the so-called free recall assumption.²²

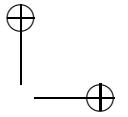
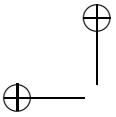
Consider first the choice between enjoying some known utility u and incurring the search cost to find out about the utility of a single alternative u_1 . If the consumer searches, alternative 1 will be preferred only if $u_1 > u$, and the consumer will otherwise enjoy utility u from the free recall assumption. It is therefore optimal to search if and only if

$$E(\max\{u_1 - u, 0\}) = \int_u^{+\infty} (x - u)dG(x) > s, \quad (8.5)$$

(throughout the chapter, our tie-breaking rule is that a consumer does not search when indifferent). The left-hand side of the inequality is zero for u larger than the maximum of the support of u_1 and it is strictly decreasing in u , from $+\infty$ to zero, for u less than the

²² See Kohn and Shavell (1974) for a formal treatment of an analogous problem.





maximum of u_1 (the derivative with respect to u is $G(u) - 1 < 0$). Hence, if \hat{u} is the value of u at which (8.5) holds with equality, then, the consumer should search alternative 1 if and only if $u < \hat{u}$.

Now assume, without loss of generality, that the consumer searches through the n alternatives from option n to option 1 in decreasing order of the index i . If, after searching all alternatives up to alternative 2, the consumer holds utility u as her best option, then she searches firm 1 if and only if $u < \hat{u}$. Now consider the decision whether or not to search alternative $i > 1$ if the best utility held thus far is u , and assume that it is optimal to search $i + 1$ if and only if $\max\{u, u_i\} < \hat{u}$. If $u \geq \hat{u}$, the consumer anticipates she will not search beyond alternative i and, by construction, it is not optimal for her to sample alternative i alone. Next, if $u < \hat{u}$, it would be desirable for the consumer to search alternative i even if she anticipates she would stop there and, if she does search beyond alternative i , it is because the expected utility from searching on exceeds $\max\{u, u_i\}$. The consumer should therefore sample alternative i . This shows by induction that, for any $i = 1, \dots, n$, the consumer should search alternative i if and only if the best utility she currently holds is $u < \hat{u}$. In other words, the consumer's optimal behavior is "myopic" in the sense that she always behaves as if there were only one alternative left to be sampled.

As an illustration consider again the case where the consumer has valuation v for some homogeneous product and her uncertainty concerns the price charged by firms $i = 1, \dots, n$, which play the same mixed strategy. Then $u_i = v - p_i$ and, letting $r = v - \hat{u}$ be the consumer's reservation price, from the definition of \hat{u} , r must solve $E(\max\{r - p_i, 0\}) = s$. If, as is typically the case, firms never charge a price larger than r in equilibrium so that $r - p_i \geq 0$ with probability 1, then r solves $r - Ep_i = s$.

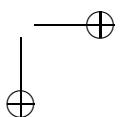
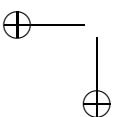
Finally, it is readily seen that the same search rule applies if there are infinitely many alternatives. The consumer then solves a dynamic programming problem where the state is the best utility uncovered thus far, u . If u is such that the consumer prefers enjoying u to searching on, then she knows that if she searches one more alternative she will stop searching afterwards. Hence, u is such that the consumer would not wish to search one more alternative so that $u \geq \hat{u}$. Furthermore, if $u < \hat{u}$, so the consumer would wish to search with one more alternative left, as in the finite horizon case, she wants to search all the more if she has the option of searching on beyond the next alternative.

We next embed this optimal search behavior in a model of price competition in a market for horizontally differentiated products.

3.2 Equilibrium and Comparative Statics

Suppose now there are n products sold by n firms with identical constant marginal costs c . There are m consumers per firm with unit demand, where a consumer's utility from buying product i at price p_i is $u_i = \mu\epsilon_i - p_i$: $\epsilon_i, i = 1, \dots, n$ are i.i.d. across products and consumers (the consumer index is dropped to ease notation) and $\mu > 0$ is a scaling parameter reflecting the degree of taste and product heterogeneity (the limit case $\mu = 0$ corresponding to a homogeneous product). The random terms ϵ_i have distribution function F and density f with support $[a, b]$. The outside option of not buying has utility zero. As is standard,²³ firms only know the distribution of ϵ_i . They select prices simultaneously in a first stage.

²³ See Perloff and Salop (1985) and Anderson et al. (1992).



Next, assume that consumers do not initially observe their realization of ϵ_i with the various products or the prices charged by firms. They may, however, find out about both through sequential search before making a purchase. We focus here on the case of random search, which can only be rationalized if consumers have identical expectations about all firms' pricing behavior, as in section 2 above. Furthermore, we look for an equilibrium where all firms charge the same price p^* so that, on the equilibrium path, observing one firm's price does not provide any information about the price of the remaining firms: in an equilibrium where different firms charge different prices (though consumers do not know *a priori* which firm is charging which price), the search process would involve learning because a consumer observing, say a low price with one firm, would infer that remaining firms are charging higher prices. The equilibrium concept is perfect Bayesian equilibrium and we impose the additional restriction that consumers hold passive beliefs about prices of firms remaining to be searched if they observe a deviation in price (this restriction would actually apply if we used the more restrictive sequential equilibrium concept because firms choose prices simultaneously and a firm's pricing should not signal what the firm does not know, as shown by Fudenberg and Tirole, 1991). We exclude the coordination failure equilibrium where firms are expected to charge such high prices that consumers choose not to initiate search (unless it is the only possible outcome).

The analysis of optimal search above can be applied to describe consumer behavior. Because all firms charge price p^* in equilibrium, a consumer's utility from buying firm i 's product is $u_i = \mu\epsilon_i - p^*$. Letting G be the distribution function of u_i , the reservation utility \hat{u} solves

$$\int_{\hat{u}}^{+\infty} (u_i - \hat{u}) dG(u_i) = \mu \int_{\hat{x}}^b (\epsilon_i - \hat{x}) f(\epsilon_i) d\epsilon_i = s, \quad (8.6)$$

where $\hat{x} = \frac{\hat{u} + p^*}{\mu}$. As can be seen from (8.6), the parameter \hat{x} is exogenously determined by the model's fundamentals: the search cost s , distribution F and scale parameter μ . The value $\mu\hat{x}$ is the consumer's reservation utility associated with uncovering her match with a product at cost s if the product was available for free: it may be derived graphically from the inverse demand for the product noting that, by integration by parts, $\int_{\hat{x}}^b (\mu\epsilon_i - \mu\hat{x}) f(\epsilon_i) d\epsilon_i = \mu \int_{\hat{x}}^b (1 - F(\epsilon_i)) d\epsilon_i$. The middle term in (8.6) is strictly decreasing in \hat{x} for $\hat{x} \in [-\infty, b]$, so \hat{x} is decreasing in search cost s and increasing in μ , reflecting the decrease in the incentive to search if search costs are higher or the uncertainty about the match realization is lower.

The implication of the consumer's search behavior for a firm's demand is straightforward. If some firm i charging price p_i competes with the option of searching another firm, then the consumer chooses to buy product i if and only if $\mu\epsilon_i - p_i \geq \hat{u} = \mu\hat{x} - p^*$, which happens with probability $1 - F\left(\hat{x} + \frac{p_i - p^*}{\mu}\right)$.

As was evidenced by Wolinsky (1986), the monopolistic competition version of this model provides a very tractable and elegant setting to analyze imperfect competition with consumer search: with an infinite number of sellers, no consumer ever "comes back" to a firm sampled earlier. Next we apply this setting, drawing on Anderson and Renault (1999), to investigate the economics of consumer search with product match heterogeneity. Under monopolistic competition the analysis is greatly simplified because the only competition a firm faces arises from a consumer's option to search on. Indeed, the other two potential sources of competition

are the outside option and all the other firms. However, if the consumer has found it optimal to start searching in the first place, the outside option is dominated by continuing to search. Similarly, if search has been preferred to purchasing some product in some round of search, then search will be preferred indefinitely. Hence, conditional on the consumer ever reaching firm i in her search, the probability that she buys product i is $1 - F\left(\hat{x} + \frac{p_i - p^*}{\mu}\right)$. In equilibrium, where $p_i = p^*$ for all i , the probability that a consumer searches on after any round of search is $F(\hat{x})$. Because there are m consumers per firm and search is random, the demand for firm i is given by

$$D(p_i, p^*) = m \frac{1 - F\left(\hat{x} + \frac{p_i - p^*}{\mu}\right)}{1 - F(\hat{x})}. \quad (8.7)$$

In order for this market to exist, consumers should search in the first place, so we need $\hat{u} = \hat{x} - \frac{p^*}{\mu} > 0$. Hence, the conditional probability of a purchase is less than monopoly demand $1 - F\left(\frac{p_i}{\mu}\right)$. This means that the firm is up against an alternative for the consumer, search, which is more attractive than the outside option, so the inverse demand is correspondingly lower.

If the distribution of the random utility terms ϵ_i satisfies the increasing hazard rate property (i.e., log-concavity of $1 - F(\cdot)$) then, as shown in Appendix B, profit is quasiconcave in price p^i and the equilibrium price has the following simple closed-form solution:

$$p^* = c + \frac{\mu[1 - F(\hat{x})]}{f(\hat{x})}. \quad (8.8)$$

Because \hat{x} is decreasing in search cost s and the hazard rate is $\frac{f}{1-F}$, the increasing hazard rate property (log-concavity) also implies that price increases in the search cost. It tends to marginal cost c if $s = 0$ so that $\hat{x} = b$, provided that $\frac{1-F(\hat{x})}{f(\hat{x})}$ tends to zero as \hat{x} tends to b (consumers keep on searching forever),²⁴ and it increases as s increases to a level such that $\hat{x} - \frac{p^*}{\mu} = 0$, where p^* is the monopoly price (at this point, search is no more attractive than the outside option). For larger search costs, the consumer gives up searching all together.²⁵ When this is the case, the market collapses (a finding we noted earlier for homogeneous products when search generates insufficient surplus at the monopoly price). Notice that this knife-edge tipping result is smoothed if consumers have heterogeneous opportunity costs for initiating search (or analogously, heterogeneous outside options as in De Cornière, 2016). Then, for given s , only those consumers with low enough “entry” costs start the search process. There is then another effect (in addition to the price effect) from changing s . This is a volume effect from the number of consumers participating in the market: a lower s both increases participation directly by decreasing costs of later search and indirectly through decreasing

²⁴ As discussed by Anderson and Renault (1999), the limit condition on the inverse hazard rate is related to the condition derived by Perloff and Salop (1985) in order for price to go to marginal cost as the number of firms becomes infinite.

²⁵ This analysis assumes that monopoly price is larger than μa so the value of \hat{x} at which price reaches its monopoly level is strictly above a .

price (see subsection 5.3 for more results on the impact of changes in search costs on consumer participation).

Regarding product and taste heterogeneity, captured by μ , the impact of a change is ambiguous. On the one hand, an increase in heterogeneity has a direct positive impact on the price captured by the direct inclusion of μ in the price expression. This reflects the standard increase in market power associated with more product differentiation if consumers are perfectly informed. However, as was pointed out above, more product differentiation implies a lower \hat{x} and hence induces more search. This intensifies competition and lowers prices (which is the case under the increasing hazard rate property). Intuitively, the first effect dominates if consumers are likely to sample many sellers so the situation is close to perfect information: this is the case if μ is large so consumers' incentive to find out about their match is high. By contrast, if product and taste heterogeneity is limited (μ small) then an increase in heterogeneity may induce a drop in price because the increased search activity of consumers intensifies price competition. We show in Appendix B that, for a low enough μ , if $f(a) = 0$ then price must be decreasing in μ : the gist of the argument is that for the values of μ such that \hat{x} is close to a (and these values are bounded away from zero as long as the search cost is strictly positive) price goes to infinity as \hat{x} goes to a .

What do we learn from the above analysis about how markets are impacted by the development of the Internet? Typically, this development has drastically reduced the level of search costs. The above model predicts a fall in the market price resulting from an increase in search activity leading to better consumer information. By contrast, in the homogeneous product settings we have discussed thus far, although price falls when search costs fall, the equilibrium search behavior is unaffected by changes in search costs (e.g., if search is sequential, consumers stop after one price quote).²⁶ As we explain below, this property remains valid for many search models with a homogeneous product, even if agents are heterogeneous. Because search costs are reduced, the additional information acquired by consumers is not necessarily associated with more resources devoted to search (e.g., more time spent searching). We actually show in Appendix B that, with the increasing hazard rate property for match values, the total search cost incurred by consumers goes down. This result, however, is obtained while keeping the supply side unaffected, in particular regarding the nature and diversity of products available in the market. We return to the issue of endogenous product choice in subsection 4.3. To capture the impact of an increased product variety due to the increased number of sellers, we now briefly discuss the oligopoly setting.

The analysis of (symmetric) oligopoly is a little more complex, but retains the qualitative properties of the monopolistic competition model.²⁷ The key difference from monopolistic competition is that some consumers reach the end of the sample set of firms, and "come back" to the one they like most or do not buy at all. This (rather appealing) feature implies that the demand facing a firm comprises two components, from those consumers passing through for the first time, and from those who have not found a good enough match to stop without sampling the full set. The latter behave like consumers with full information in a standard

²⁶ With simultaneous search, the probability that consumers sample a second price increases only in the high search intensity equilibrium. It actually decreases in the low search intensity equilibrium. In any case the maximum number of prices sampled remains two.

²⁷ Anderson and Renault (1999) show that the comparative statics results described above for monopolistic competition still hold with the increasing hazard rate property and a covered market.

product differentiation context, subject to the proviso that their valuations for all goods (in equilibrium) are below the stop threshold.

The additional driver of the equilibrium price is the competition for those who may come back. The more firms there are, the fiercer the competition (as per standard discrete choice models), because a firm has to beat the best of n other options including the outside option. Hence, the price is higher when there are fewer firms. Put another way, individual demand is more inelastic with fewer firms because there are fewer substitutes. Anderson and Renault (1999) show this intuition is borne out in a covered market under the increasing hazard rate assumption.²⁸ They also consider endogenous entry and find that search costs exacerbate the over-entry that characterizes such markets with perfect consumer information:²⁹ higher search costs make entry more profitable by increasing market power, and they reduce the benefit of having more variety because consumers sample fewer products. The reduced search costs resulting from online shopping should be expected to mitigate this inefficiency. Furthermore, it is likely that the online technology reduces entry costs so that the number of sellers in the market increases despite the weaker market power. This increased product variety is desirable. There is also a market expansion effect due to increased buyer participation when consumers are heterogeneous regarding the decision to start search, as discussed above and further in subsection 5.3. The welfare implications of this market expansion are non-trivial: if the volume effect dominates the price effect, then profitability of entry increases. Moreover, as we explain in subsection 5.3, if search costs are heterogeneous the price effect might also be positive.

Finally, Zhou (2014) provides some caveats to these conclusions. He considers a multiproduct search market in which consumers search for more than one good, and that a search reveals the match values (and prices) of all goods sold in the store. For simplicity, suppose that valuations for goods are independent, so they are neither complements nor substitutes, *per se*. Notice that the assumption of free recall implies that a consumer will never buy one product and keep searching for others, for there is a chance of finding a better product later, and she can always come back costlessly. This feature implies that reducing the price of one product at a firm will also raise the demand for the firm's other products, which Zhou calls the *joint search effect*: a firm can thus induce consumers to stop and buy its other products, rendering products complements. If search costs were to decrease, the joint search effect is weakened and, as he shows, this can dominate the standard effect that works through prolonging search, and actually cause prices to rise.

3.3 Mergers and Cartels

Moraga-González and Petrikaitė (2013) study the impact of a merger in an oligopoly with consumer search and differentiated products. There are two main insights. First, anticipating that insider firms will charge a higher price than outsiders, consumers begin searching through the latter first.³⁰ This order of search penalizes the merged entity by reducing the consumer base to which it has access as compared to the pre-merger situation where consumers search

²⁸ We are not aware of any general results when allowing for an outside option.

²⁹ See Anderson, De Palma, and Nesterov (1995) for the perfect information case.

³⁰ Ordered search is discussed in considerably more detail in the next subsection and all the sections' final subsections.

randomly. If search costs are large enough, this adverse effect may actually outweigh the benefit from coordinated pricing afforded by the merger, so the merged entity's profit ends up being lower than the participating firms' pre-merger joint profit. This result contrasts with previous literature on mergers with price competition. This type of "merger paradox" is typically expected to arise in a context of quantity rather than price competition. Here the disincentive to merge results from the impact of the merger on consumer search behavior rather than from its impact on the strategic interaction between firms.

The second insight is that the above may only be a short-run consequence of the merger. In the long run, the merged entity might be able to reorganize its commercial activities so as to facilitate the consumers' access to information about its products and prices. When searching that firm, consumers can then obtain information about their match with a wide range of products by incurring a low search cost. This may then induce consumers to search through the insider products before the outsider products. This prominence of insider products may lead the merged firm to charge lower prices than outsiders. Because outsiders are searched last, their consumer base shrinks and for high enough search costs their profit is below its pre-merger level. Finally, thanks to search cost economies, consumer welfare as well as social welfare may be increased by the merger.

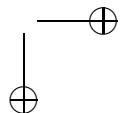
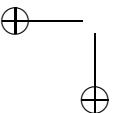
These results underscore that allowing for search costs brings new considerations to the analysis of mergers. Much of the standard merger analysis contrasts the anti-competitive effect of mergers with the pro-competitive benefits due to the exploitation of economies of scale by the merged firm. Here the long-term benefit of the merger stems from the search costs saved by consumers, who may find out about all the products sold by the merged firm while incurring the search cost once. The results also show that the prediction about the impact of the merger on prices may critically depend on how the consumers' search behavior is affected.

Janssen and Moraga-González (2008) explore the impact of mergers in a market for a homogeneous product where consumers have heterogeneous search costs. Their setting is similar to that of Stahl (1989), described in subsection 5.2: some of their findings confirm the results in Moraga-González and Petrikaitė (2013). In particular, they also find a merger paradox and the possibility that a merger with no efficiency gain in production may benefit consumers, although the underlying mechanisms are quite different.

Petrikaitė (2016) looks at the factors facilitating collusion in such a setting. The main insight is that a search cost increase has two countervailing effects on cartel stability.³¹ On the one hand, it decreases the deviation gain because fewer consumers are aware of the deviation. On the other hand, it makes the punishment milder because a higher search cost means more market power.³² In total though, she shows that the first effect dominates and cartel stability is enhanced if search costs are larger.

³¹ The working definition of cartel stability here is based on the critical discount factor analysis from simple trigger strategy equilibria in infinitely repeated games. Full collusion is sustained as long as the discount factor, $\delta \geq \hat{\delta} = \frac{\pi^D - \pi^C}{\pi^D - \pi^N}$ where the superscripts denote deviation, cartel (or collusion), and Nash (i.e., punishment) payoffs. Greater stability is interpreted as lower $\hat{\delta}$, which is achieved if the excess of deviation over collusive profits goes down relative to deviation over punishment. The condition can be phrased alternatively in terms of differences over punishment profits as $\hat{\delta} = 1 - \frac{\pi^C - \pi^N}{\pi^D - \pi^N}$.

³² These effects are quite analogous to increasing product differentiation in a perfect information setting.



3.4 Ordered Search

In keeping with the corresponding literature, we have treated search as random thus far (with the exception of the merger analysis of the previous subsection). Even when search is sequential, we have assumed that the order of search is chosen randomly across options. This has the convenient property of allowing for symmetric equilibria, but there are natural circumstances when sequential search follows an order (which might sometimes differ across consumers). For example, consumers might travel to closer stores before those further away, so geography intervenes (see, for example, the description in section 5.4 of the model of Arbatskaya, 2007). More subtly, there may be equilibria at which consumers are expected to follow a particular order, and firms price accordingly. Indeed, as argued by Armstrong (2016), the symmetric equilibrium (of Wolinsky, 1986, and Anderson and Renault, 1999) is unstable in the sense that if more consumers choose a particular firm first, then, insofar as that firm's price is lower (which is a key point that we develop below), then other consumers would also want to search the more popular firm first. This tips the equilibrium naturally to one of ordered search. If, in addition, being searched earlier is more profitable, this introduces the possibility that firms might pay for more prominent positions on websites (as with position auctions), or on supermarket shelves (as with slotting allowances), or closer to consumers (as with geographical locations). Such paying for prominence (in the terminology of Armstrong and Zhou, 2011) may involve a variety of firm practices as discussed in Armstrong (2016), such as advertising a low price (we briefly return to price-directed search in the conclusion) or non-price advertising where consumers coordinate their search on the seller that advertises the most, as in Haan and Moraga-González (2011).

Taking the leading example of the differentiated products model of section 3.2 (match values are heterogeneous, but i.i.d., as per Wolinsky, 1986) we suppose now that prices are expected to differ across firms. Then the consumer will clearly follow the order of increasing expected prices, although we still need to determine her stopping decision. We argued above in section 3.2 that if all prices are expected to be the same, then she searches on if the best match thus far is less than some threshold; moreover, this is true independent of the number of firms left. That is, she uses a myopic rule, and additional equally attractive search options do not make search more appealing. As should be apparent intuitively, this logic extends to when prices differ. If firm i has the next lowest expected price, and that price is p_i^* , then the consumer searches if her best utility with firms she has already visited is less than p_i^* . The prospects from searching firm i cannot improve when the later options are increasingly less attractive.

This case already covers most of the models in the literature on ordered search with product heterogeneity. Because consumers search by lowest prices first, consistency requires that firms searched earlier do indeed want to price lower. To see the tension, consider a simple duopoly case. Let us compare the two firms' price elasticity of demand if they charge the same price. Insofar as the first firm sampled gets all traffic (the first bite of the apple), then its demand is larger, which all other things equal implies a lower price elasticity. However, as we now argue, its demand derivative is larger (in absolute terms), and hence the source of the potential ambiguity. It is instructive into the nature of pricing to see why demand derivatives are different by considering the simpler case of a covered market. Then, in the absence of search the derivatives are the same in a duopoly (each firm loses the marginal consumers to its rival from a price hike). With firm 1 being visited first, we can split the consumers into two

groups. One group buys directly from firm 1; it stops before finding out its match with firm 2. The rest of the consumers sample both firms, and are thus fully informed. All consumers observe a price rise by firm 1, and so it loses consumers to firm 2 from both the fully informed and those who erstwhile stopped at it. But only searchers observe an (unanticipated) price hike from firm 2. That is, firm 2 does not lose consumers over the transom between it and its rival at the margin of discouraging them from searching, whereas firm 1 does. Therefore 1's demand derivative is larger, as claimed.

Extending the reasoning to several firms, the demand of a firm searched later is smaller, but is less sensitive to price changes because the firm cannot communicate a price drop: consumer decisions to stop are driven by (rationally) expected prices, but not actual ones. The earlier is a firm in the search order, the more consumers see a price change, but the larger is demand. However, as shown in Armstrong (2016), an increasing hazard rate in the product match distribution (log-concavity) is a sufficient condition for demand to be more elastic, and hence for price to be lower earlier in the search order.³³ When a consumer arrives at a later firm, the firm can infer that she does not like the earlier products: she searches even if she expects later prices to be higher. This gives the later firms extra market power.

The possibility of bidding for prominence adds yet a further requirement on the equilibrium. In addition to prices rising with the consumer search order, it should also be the case that profits fall. Otherwise, if this condition were revoked at some point in the order, firms would want to wait for later positions, and not pay a premium to be earlier. The requirement is actually necessarily satisfied from a simple revealed preference argument. Indeed, a firm can always choose to charge the equilibrium price of its successor in the queue. In this symmetric product setting, this necessarily yields a larger profit for the earlier firm. Hence if it chooses to price lower, it must be earning more profit than the next firm.

Unfortunately, for reasons similar to those we discussed in the case of random search, there is no existence argument relying on general properties on the match distribution (such as an increasing hazard rate). The early articles on the topic, such as Armstrong, Vickers, and Zhou (2009a) or Zhou (2011), used a uniform match distribution. As with random search, the monopolistic competition setting provides a simple fix. It is then easy to characterize a symmetric price equilibrium. Because price is expected to be the same at all firms, a consumer's search behavior is identical to that derived for random search. For reasons analogous to those discussed in subsection 3.2, the infinite number of firms implies that each firm only competes with the consumer's option to continue search. Because the consumer's search prospects are stationary, all firms charge the same price: although firms placed earlier sell more, their demand derivative is also larger by exactly the same proportion, so elasticity is the same at all positions. As with random search, the equilibrium price is given by (8.8). In this case, then consumer welfare and aggregate firm profits remain unaltered, so market performance is neutral. However, the first firm searched has more equilibrium sales than the second, etc., so that profits decrease with exponential decay through the order (all prices are the same, and each firm after the first one serves a fraction $F(\hat{x})$ of its predecessor's demand). This opens up the possibility that firms would want to pay for prominence, which we reconsider in the next section where firms are assumed to differ *ex ante*.

³³ Rhodes and Zhou (2016) obtain a similar result in a variant with two firms, one of which is multiproduct.

4 SELLER HETEROGENEITY

Much attention has been devoted to the theoretical possibility of price dispersion that results solely from imperfect consumer information (see section 2). In practice, it is likely that sellers that charge different prices face different costs and/or demand.

4.1 Searching for an Efficient Seller

A (somewhat obvious) first point is that introducing efficiency differences across sellers naturally generates price dispersion. Following Reinganum (1979), assume a market where m consumers have identical individual downward-sloping demands, $d(p)$ at price p , for a homogeneous product sold by a continuum of firms. Let firm i have constant marginal costs, c_i , and these differ across firms according to a continuous distribution on an interval $[\underline{c}, \bar{c}]$, with $\underline{c} > 0$. Assume $d(p)$ is well behaved so that for all i , firm i 's monopoly profit $md(p)(p - c_i)$ is single peaked in p , and the corresponding monopoly price is $p^m(c_i)$. A well-behaved demand also ensures that monopoly price is strictly increasing in marginal cost. Consumers can only observe prices through random sequential search, where $s > 0$ is the search cost.³⁴ The marginal cost distribution is common knowledge but consumers do not know its realization for each firm.

Consider now the following strategies. A consumer searches at least one firm and keeps on searching as long as the best price observed so far strictly exceeds some reservation price r . Firm i charges $p^m(c_i)$ if and only if $p^m(c_i) \leq r$, and charges r otherwise. Let \tilde{c} be the marginal cost such that $p^m(\tilde{c}) = r$ (which exists as long as r is not too large). Optimal search requires that r solves $r - E_{c_i < \tilde{c}} p^m(c_i) = s$ (note that, due to the continuum of firms assumption, a consumer draws from the same price distribution at any round of search). As r increases from $p^m(\underline{c})$ to $p^m(\bar{c})$, the left-hand side increases from 0 to $p^m(\bar{c}) - Ep^m(c)$. For s in this range, r defines an optimal strategy for consumers given the firms' pricing (as long as consumer surplus at all those monopoly prices exceeds s). The proposed strategies for firms are profit maximizing because all prices are below r , so that all consumers buy from the first firm sampled. Thus each firm should maximize its profit on the mass of consumers $\frac{m}{n}$ visiting it first, subject to the constraint that it should not price above r . For $s > p^m(\bar{c}) - Ep^m(c_i)$, the equilibrium entails all firms charging their monopoly prices.

Although introducing cost heterogeneity generates price dispersion that makes search potentially appealing, it does not substantially overturn the Diamond paradox. Only the least efficient firms are forced to price below their monopoly prices. Furthermore, as the search cost tends to zero, the equilibrium prices all tend to the lowest monopoly price, rather than marginal cost (which might have been expected in a market where consumer information about prices is almost free). Note that inefficient firms may remain active even if the search cost is very small. Regarding consumer search behavior, there is no search beyond the first firm in equilibrium, no matter how small the search cost.

We next return to the monopolistic competition model with heterogeneous products of subsection 3.2, assuming now that each firm's marginal cost is distributed identically and independently across firms on support $[\underline{c}, \bar{c}]$ with distribution function H . Consumers do not know firm i 's marginal cost, but the cost distribution is common knowledge. Suppose that if

³⁴ Bénabou (1993) and Bénabou and Gertner (1993) allow for heterogeneity in both production and search costs.

firm i has cost c_i , it is expected to charge price $p(c_i)$.³⁵ Then a consumer's utility from buying product i is the random variable $u_i = \epsilon_i - p(c_i)$ (we here assume $\mu = 1$ to simplify notation). The analysis of subsection 3.1 applies, so we may define the threshold utility \hat{u} above which a consumer stops searching. In equilibrium, a firm's pricing decision depends on this threshold, so we denote the equilibrium price of a firm with cost c_i by $p^*(c_i, \hat{u})$. Again letting G denote the distribution function of u_i , \hat{u} solves

$$\int_{\hat{u}}^{+\infty} (u_i - \hat{u}) dG(u_i) = \int_{\underline{c}}^{\bar{c}} \int_{\hat{u}+p^*(c_i, \hat{u})}^b (\epsilon_i - p^*(c_i, \hat{u}) - \hat{u}) dF(\epsilon_i) dH(c_i) = s. \quad (8.9)$$

As shown in Appendix B, profit maximization implies

$$p^*(c_i, \hat{u}) = c_i + \frac{1 - F(p^*(c_i, \hat{u}) + \hat{u})}{f(p^*(c_i, \hat{u}) + \hat{u})}. \quad (8.10)$$

Appendix B also shows that the increasing hazard rate property for the match distribution ensures that a firm reacts to a higher reservation utility by charging a lower price, although it adjusts its price by an amount that is less than the reservation utility change. It follows that $p^*(c_i, \hat{u}) + \hat{u}$ is strictly increasing in \hat{u} . This is because log-concavity of demand (implied by the increasing hazard rate) induces a pass-through of demand or cost shifts less than unity (Anderson and Renault, 2003).³⁶ Hence the middle term in (8.9) is strictly decreasing in \hat{u} (for each realization of c_i the integral with respect to ϵ_i is strictly decreasing). It follows that \hat{u} is uniquely defined and strictly decreasing in search cost s : an increased search cost makes search less attractive.

The above analysis shows that the predictions obtained with homogeneous firms regarding the impact of search costs readily extend to a setting where firms selling differentiated products have different marginal costs: lower search costs induce more search and hence lower prices. However, this setting provides new insights for very high or very low search costs. Because high-cost firms charge higher prices, the critical match value at which consumers prefer searching to buying their product, $p(c_i, \hat{u}) + \hat{u}$, reaches b (as s decreases) earlier for high-cost (less efficient) firms than for low-cost (more efficient) firms. Hence, less efficient firms become inactive if search costs are sufficiently low.³⁷ For very high search costs, however, efficient firms exert a positive externality on (and suffer a negative externality from) inefficient firms. This is because the search cost level at which \hat{u} reaches zero (at which point the market collapses) depends on the entire distribution of marginal costs. Hence, this critical search cost value would be higher if the market comprised only efficient firms, and lower if consumers anticipated that all firms are inefficient.

Finally, allowing for cost heterogeneity provides some insight into the impact of search costs on price dispersion. With homogeneous products, dispersion is reduced with lower search costs to the extent that more firms charge the reservation price. By contrast, with horizontal product differentiation, as long as the reduction in search costs has no impact on the

³⁵ This framework is a simplified version of that analyzed by Bar-Isaac, Caruana, and Cunat (2012) to which we return in subsection 4.3.

³⁶ Weyl and Fabinger (2013) provide further results on pass-through in oligopoly.

³⁷ If these inefficient firms drop out when they can no longer sell, the value of \hat{u} is affected because the upper bound on marginal costs is endogenously decreasing in \hat{u} .

mix of firms (in terms of marginal costs) participating in the market, there is no clear reason why price dispersion should be reduced: whether inefficient firms lower their prices more than low-cost firms in response to an increase in \hat{u} depends on properties of the match value distribution, beyond the increasing hazard rate restriction.³⁸ However, if search costs drop so much as to drive inefficient firms from the market, then there is a clear pressure towards reducing the range of prices in the market.

4.2 Learning about Sellers' Efficiency

Although it is convenient to assume that firm characteristics are idiosyncratic, doing so leaves out an important feature of consumer search, which concerns how consumers can learn about market conditions. Unfortunately, accounting for such learning complicates the analysis considerably. Here we describe the main insights that can be derived from simple specifications and also point out the added difficulties associated with this type of analysis.³⁹

Consider again a market for a homogeneous product where sellers have different marginal costs. By contrast with our previous analysis, assume now that these costs are positively correlated due to common shocks such as changes in input prices. As in Bénabou and Gertner (1993), think of a simple duopoly model where consumers first observe one price and then decide whether or not to incur a search cost to find out about the other price. Provided that price is monotonically increasing in cost, a high price observed by the consumer in the first round leads her to revise her expectations upwards regarding the price she will find at the second firm. She will revise her price expectations unfavorably all the more if the uncertainty regarding the common shock is large relative to the uncertainty regarding the idiosyncratic cost differences among firms. This is because more variability in the common cost component means a higher correlation between the firms' costs. This per se provides an incentive for firms to charge higher prices when global cost uncertainty increases. However, as emphasized by Bénabou and Gertner (1993), an increased uncertainty and an increased correlation in costs also impact the consumers' incentives to search, which in turn affects the firms' pricing. In particular, increased cost uncertainty resulting in more variability in prices makes search more attractive, which induces more competition and pushes prices downwards. Bénabou and Gertner find that a higher global cost uncertainty increases prices if search costs are high but has the reverse effect otherwise.

Note that if there is no idiosyncratic heterogeneity in firm costs and firms play a symmetric pure strategy equilibrium, consumers learn about the common shock perfectly with only one price observation. This would be the case, for instance, in the model with product and taste heterogeneity of section 3 if there were some cost uncertainty although all firms would have the same cost. Indeed, Janssen and Shelegia (2016), who consider a duopoly version of the Wolinsky (1986) setting, find that the equilibrium price is higher when consumers are uncertain about marginal cost than if there is no such uncertainty. They find, however, that, for a high enough search cost, price with cost uncertainty is decreasing in the search cost, whereas the price with no uncertainty is still increasing, and both prices become equal

³⁸ For a uniform taste distribution (i.e., linear demand), the equilibrium price is $p^*(c_i, \hat{u}) = \frac{1}{2}(1 + c_i - \hat{u})$ so that all prices respond the same way to a change in \hat{u} .

³⁹ Our presentation leaves out a fairly substantial literature that looks at learning and price dynamics in search models such as Dana (1994), Fershtman and Fishman (1992), Fishman (1996) or Tappata (2009). These studies typically assume homogeneous products and simultaneous search.

to the monopoly price if the search cost is high enough so consumers stop searching all together. Janssen, Pichler, and Weidenholzer (2011) also explore the role of common cost uncertainty in a market for a homogeneous product with symmetric firms. In their setting, there is, however, price dispersion due to mixed strategies resulting from some heterogeneity in consumer information.⁴⁰ Because of the mixed pricing strategies, consumers cannot learn about the cost shock perfectly. Still, they find the expected price is larger with cost uncertainty.⁴¹

As emphasized by Janssen and Shelegia (2016), when comparing the cost uncertainty setting with the standard setting where marginal cost is common knowledge, we should be cautious about what we assume about consumer beliefs. Cost uncertainty naturally leads us to assume that, in a pure strategy equilibrium, any price that is in the support of equilibrium prices is interpreted as a sufficient statistic for the marginal cost. Thus, when a firm deviates upwards, it expects consumers to infer from the new price that the other firm is also charging a higher price. By contrast, the standard approach with no cost uncertainty presented in section 3 assumes consumers have passive beliefs. However, if consumers expected the slightest positive correlation in marginal costs, the justification for passive beliefs would no longer be valid. Janssen and Shelegia show that the equilibrium behavior with cost uncertainty is analogous to the equilibrium behavior in the standard model where consumers have “symmetric” rather than passive beliefs, so they always expect the price at the second firm to equal the price at the first firm they visit.

An important methodological remark is in order here. Once there is learning, the standard optimal search analysis described in subsection 3.1 no longer applies. First, even if the optimal search behavior involves a reservation utility, it should evolve over the search process and depend on the search history because learning affects a consumer’s expectations. Second, as first noted by Rothschild (1974), an even more drastic implication of learning is that the optimal rule may not be based on a reservation utility level.⁴²

We next give a brief discussion of endogenous choice of product design.

4.3 Product Choice and Search

In our discussion of the impact of the search cost reduction associated with the development of Internet search, we pointed out that it is critical to take into account supply-side adjustments, in particular regarding the types of products that are available for sale. In the search framework with horizontal differentiation and heterogeneous costs of subsection 4.1, a drop in search costs shifts a firm’s inverse demand downwards, due to the improved attractiveness of search reflected in a higher \hat{u} . Firms might then want to adjust their product designs appropriately. The analysis in Johnson and Myatt (2006) provides a simple setting through which this choice can be investigated.

Here we borrow from Bar-Isaac et al. (2012) to understand the change in product mix that is induced by lower search costs. Johnson and Myatt (2006) describe a firm’s choice whether to make its product widely appealing to a large customer base or instead very appealing to

⁴⁰ They use the setting of Stahl (1989) discussed below in subsection 5.2.

⁴¹ Note though that the equilibrium they characterize only exists for certain parameter configurations. In particular, the search cost should be large enough relative to the uncertainty in marginal production costs.

⁴² See Janssen, Parakhonyak, and Parakhonyak (2016) for an equilibrium search analysis without reservation price search rules.

a small niche of customers who are willing to pay a lot for it. This choice can be depicted as the firm choosing to rotate its inverse demand by making it flatter or taller (with a flatter curve corresponding to a larger customer base). They show that the optimal choice is always extreme, so a firm chooses either the flattest or tallest inverse demand. Whether a firm chooses flat or tall depends on whether its marginal cost is high or low relative to the vertical position of the inverse demand that reflects how much buyers value the product “on average.” If marginal cost is low, then it is optimal for the firm to sell a lot and it will select the mass market product design.⁴³ If, on the contrary, marginal cost is high, the firm prefers to sell less and selects the alternative niche market product design. Then, this analysis leads us to expect that there is some threshold cost such that all firms with marginal cost above the threshold select a niche product, whereas firms with marginal cost below the threshold select a mass market product. Bar-Isaac et al. (2012) characterize such an equilibrium in an environment where firms differ by product quality (where a high quality is equivalent to a low marginal cost in our setting). Now if the search cost decreases, leading to an increase in \hat{u} , each buyer’s willingness to pay for the product (as opposed to searching on) drops, making the niche product strategy more attractive. As a result the product mix involves more niche products.⁴⁴

In terms of our analysis in section 3, choosing a niche product corresponds to choosing a product with a large μ , reflecting a larger heterogeneity in match values. Because such products are more attractive to search (\hat{e} being increasing in μ), the change in the product mix increases search activity beyond the level that would ensue if product design was exogenous. This is well reflected in the approach taken by Larson (2013). In accord with the results noted above, he shows that firms’ equilibrium choices are extreme. He emphasizes the feedback loop in the equilibrium: high reservation values encourage search and so encourage firms to choose highly differentiated (niche) products, which in turn further encourages search. Likewise, low reservation values encourage low differentiation. Larson (2013) also argues that common specifications of consumer preferences are likely to give an asymmetric equilibrium with niche and generic products coexisting.

4.4 Ordered Search

Thus far in this section, although we have allowed firms to differ in marginal costs and product quality, we have postulated that consumers search randomly. This requires that the characteristics of a firm remain unknown to a consumer until she visits it. Further assuming an infinite number of sellers allows for describing the optimal search rule as a simple stationary stopping rule. Yet, if firms differ in costs or quality, we would expect them to try to make this information known if it is favorable. A central theme here is that paying for prominence is a means of credibly transmitting such good news. We first reconsider the optimal search problem from a more general perspective than the one of subsection 3.4.

Weitzman (1979) delivers a very clean characterization of optimal search behavior. Suppose that options can differ by the distribution of the utility that they are expected to yield, so that we write G_i as the utility distribution for option i . We retain the independence of the distributions, but we can dispense with their being identical. We also retain costless recall,

⁴³ These results recall those of Lewis and Sappington (1994).

⁴⁴ Bar-Isaac et al. (2012) show that the same predictions can arise in a setting with *ex ante* homogeneous firms, where firms mix between the two product designs.

and search cost s per search.⁴⁵ As Weitzman (1979) shows, the solution to the consumer's problem has a simple algorithmic condition that determines what to do. Surprisingly perhaps, the solution is not to search in order of highest expected value (although this is true in the simplest cases), but is not much more complex: Weitzman's result is gratifyingly simple.

Define for each option a score value \hat{u}_i by

$$\int_{\hat{u}_i}^{+\infty} (u_i - \hat{u}_i) dG_i(u_i) = s, \quad (8.11)$$

which generalizes (8.6). The consumer's optimal search is to go through these scores in decreasing order until the next option delivers a score below what is currently held: this may involve costless recall, which is to go back and select whatever previously held option is best. These scores are myopic reservation values, where \hat{u}_i is the highest utility currently held such that searching option i is optimal. The rule can readily be proved by backward induction similarly to the argument in subsection 3.1.⁴⁶ There are important cases where the reservation values are ranked in an unambiguous manner. First, a distribution has a higher score if options are ordered in terms of first-order stochastic dominance, as should be expected. One simple case is when match distributions are i.i.d. and only prices differ. Second, a wider spread yields a higher score when all options are mean preserving spreads of one another: this reflects that it is more appealing to search riskier options. In line with the previous section, this means that consumers would choose to search sellers of niche products first (*ceteris paribus*).

For the most part, the existing literature on ordered search with firm heterogeneity has considered products with different qualities, so that there is a first-order stochastic dominance ordering of match value distributions. Even in this simple case, the characterization of a price equilibrium is quite challenging. To see why, consider again the monopolistic competition version of the model in Wolinsky (1986), which has proved in many instances to be quite tractable. Suppose, as in Bar-Isaac et al. (2012) that match values for product i are given by $v_i + \epsilon_i$, where ϵ_i is the idiosyncratic match utility that differs across buyers, whereas v_i is some quality measure that is evaluated in the same manner by all. In contrast with the analysis in subsections 4.1 or 4.3, consumers know v_i and may therefore devise an optimal search order, which will be the same for all because they are *ex ante* identical. The optimal order does not necessarily follow the order of v_i because it also takes into account the prices expected by consumers at each firm, p_i for firm i . With no loss of generality, assume the reservation values defined by (8.11) for each firm i decrease in index i , $\hat{u}_1 > \hat{u}_2 > \hat{u}_3 \dots$. Then consumers start search with firm 1 and move on to firm 2 and so on as long as they hold a utility that is less than the next reservation value. Hence, a consumer searches firm 2 if $\epsilon_1 < \hat{u}_2 - v_1 + p_1$. Among such consumers, those whose match with firm 1 is $\epsilon_1 > \hat{u}_3 - v_1 + p_1$ (and there are such consumers because $\hat{u}_3 < \hat{u}_2$) will not want to search firm 3, and will pick whichever firm they like better between firms 1 and 2. Hence, there are consumers coming back despite the infinite number of firms. This precludes the simple existence arguments based on the increasing hazard rate property we have used thus far for monopolistic competition. Besides, the characterization of equilibrium pricing must deal with an infinite number of demand

⁴⁵ Weitzman (1979) does allow for different costs per search, as well as discounting of options searched later, with the same characterization we give below.

⁴⁶ See Armstrong (2016) for an alternative proof of the result and for some more discussion of Weitzman's rule.

terms reflecting the populations of returning consumers after additional rounds of search, compounded with the asymmetries resulting from the search order and the heterogeneity in product qualities.

Armstrong, Vickers, and Zhou (2009a), in their extension where firms have different qualities,⁴⁷ overcome these obstacles by supposing that only one firm obtains a prominent position and then the other firms (an infinite number) are searched randomly. Then a consumer who decides to search after visiting the prominent firm enters a world like that described in subsection 4.1 so she keeps on searching until she obtains a utility above her reservation value, which is now stationary. Then no consumer ever returns to any firm. They show that the firm with the highest quality has the highest willingness to pay for prominence. Another route is to simplify the model by only considering a duopoly as in Song (2017).⁴⁸ Still the analysis is quite involved even when using uniform distributions of matches.

An alternative is to follow Chen and He (2011) or Athey and Ellison (2011) and assume that all prices are exogenous and identical.⁴⁹ They suppose that firms only differ in the probability that a consumer is “matched” with the product, i.e., the probability that the consumer wants to buy. Let β_i denote that probability for firm i then, normalizing price and the consumer population to 1, a firm’s profit is merely β_i times the probability that it is visited. Taking a two-firm example and assuming firm 1 is visited first, firm 1 earns β_1 and firm 2 earns $(1 - \beta_1)\beta_2$. Firm 2 is willing to pay $\beta_1\beta_2$ to be visited first rather than second. This is symmetric and therefore both firms have the same incremental value from being ranked first. This means that consumers have no reason to expect that the firm ranked first (say on a web page) is the most attractive and that they should therefore start with that firm. One solution is to assume that firms only know their own probability. For instance, if probabilities are drawn from i.i.d. distributions with mean λ , then firm i ’s incremental value for being ahead is $\beta_i - (1 - \lambda)\beta_i = \lambda\beta_i$, which is increasing in probability β_i . As explained in subsection 5.4, different search costs for different consumers can also fix this problem.

Anderson and Renault (2016) introduce a demand system that captures some key features of pricing with ordered search while allowing for multiple dimensions of *ex ante* product heterogeneity. As in Athey and Ellison (2011) or Chen and He (2011), consumers stop search upon finding a product with which they are matched. However, this behavior is not exogenously assumed but here results from the firms’ pricing decisions. This is because the demand function ensures that, conditional on being matched, a consumer has a high enough willingness to pay for the product that the firm always finds it profitable to charge a price low enough that she does not search. Firm pricing has the key property that firms price in such a way that consumers find it optimal to follow whatever search order is assumed. The demand system allows for three dimensions of heterogeneity. First, as in the simple fixed price setting above, products differ in how popular they are (the probability of matching a consumer’s need). Second, they differ in how valuable they are to consumers, which is captured by the minimum valuation of a consumer who is matched with the product. Finally, they differ in the heterogeneity of tastes captured by a measure of the thickness of the upper tail of the match

⁴⁷ Their specification of quality is different from the one we have used above.

⁴⁸ Song (2017), more generally, supposes that one match distribution is a mean-preserving spread of the other, making it more appealing to search. He shows that joint profits are maximized if consumers search first the more spread distribution.

⁴⁹ Chen and He (2011) assume that all products that a given consumer cares about are perfect substitutes, so the Diamond paradox ensues and all firms charge the monopoly price endogenously.

distribution. The latter engages the characterization of the optimal search rule in Weitzman (1979) in quite a general manner deeper than simply first-order stochastic dominance (as was previously considered). One insight into quality signaling by paying for prominence is that, with endogenous pricing, a firm selling a less popular product may be willing to pay more to be ahead in the search order.

5 BUYER HETEROGENEITY

Buyer heterogeneity has been a much more active and fruitful research direction than seller heterogeneity. More specifically, search cost heterogeneity is a dimension of the search equilibrium problem that has delivered significant progress in our understanding of the impact of search costs on competition. We start though by discussing demand heterogeneity.

5.1 Demand Heterogeneity with a Homogeneous Product

It might be expected that demand heterogeneity could overcome the Diamond paradox. Some firms could specialize in selling small amounts to consumers with high valuations, while others could charge low prices and cater to more people, including some consumers with lower valuations. We now explain why this is not a very promising research agenda for looking at the impact of reduced search costs.

First, following Stiglitz (1979), it is fairly well known that, if the first search is costly and consumers have (possibly heterogeneous) unit demands, then the market necessarily unravels if consumers with identical search costs must search sequentially to find out about prices.⁵⁰ To recapitulate the argument, let us again denote the lowest candidate equilibrium price by \underline{p} . A firm charging that price knows that all consumers entering the market necessarily have a valuation of at least $\underline{p} + s$, because they were willing to incur the first search cost. The firm could then deviate by charging a slightly higher price and not lose any customers. Hence there can be no such \underline{p} , and therefore no equilibrium. It is straightforward to use the same line of argument to establish that the unique equilibrium has all firms charging the monopoly price if the first search is free and monopoly profit is single peaked; moreover, no consumer searches beyond the first firm. To the best of our knowledge, it has not been investigated whether or not this result generalizes to price-sensitive demands and/or a monopoly profit that is not single peaked. However, we argue below that, although it is possible to devise situations where the monopoly price may not sustain, the resulting predictions are not substantially different from those of the Diamond paradox.

One can easily see that then there may be single-price equilibria different from the monopoly price if monopoly profit is not single peaked. Consider a local maximum of monopoly profit at some price p , such that no local maximum at a lower price yields a larger profit. Because p is a local maximum, there exists $h > 0$ such that price p maximizes profit on $[p, p + h]$. Now, if there is a finite number of consumer types and if search cost s is small enough, then a price hike with a magnitude of at least h starting from p will result in a drop in surplus larger than s for all consumers. Hence it is possible to sustain an equilibrium with

⁵⁰ The argument here does not require homogeneous products, just that consumers know their valuations before search.

n firms where each firm charges p . In such an equilibrium, no consumer searches and each firm earns one n th of the monopoly profit at price p . By construction, a downward deviation is unprofitable. Furthermore, any price increase such that consumers choose not to search must be less than h and thus does not increase profit. Then we have an equilibrium if the first price quote is for free. Furthermore, if individual demands are price elastic and s is less than the minimum consumer surplus at price p , consumers will participate in the market and the equilibrium is sustained even if the first price quote costs s . Although this shows that it is possible to obtain equilibrium pricing below the monopoly price (if that price corresponds to a local maximum above p), the candidate equilibrium, like the Diamond outcome, involves no search and pricing that is unaffected by changes in the search cost or the number of firms.

Multiple peaks in monopoly profit also allow price dispersion in the form of a mixed strategy equilibrium. This is the case even without demand heterogeneity. Consider a single consumer type and a continuous monopoly profit. Assume there exists some local maximum p less than the monopoly price but yielding more profit than any price below p . Because p is below the monopoly price, continuity of profit requires that there exists at least one price r at which profit is increasing in price and is the same as at price p . Take the lowest such price so that no price in $[p, r]$ generates more profit than prices r and p . Letting S denote the consumer's surplus, assume now that $s < S(p) - S(r)$. Then there exists $\alpha \in [0, 1]$ such that $\alpha(S(p) - S(r)) = s$. This defines a symmetric mixed strategy equilibrium with support $[p, r]$ where firms charge price p with probability α and the consumers' search behavior is characterized by a reservation price of r . Note though that consumers do not search. Furthermore, as search cost s decreases, equilibrium prices do not change and the probability α of a low price decreases, so expected prices rise.

Demand heterogeneity is an obvious source of multiple peaks in monopoly demand even if individual demands are well behaved. More importantly, it creates new possibilities for price mixed strategies even if monopoly profit is single peaked. Suppose there are two consumer types with price-sensitive demands. Because the two types have different surplus functions, in a symmetric mixed strategy equilibrium, they will have different reservation prices, r_h and r_ℓ , with $r_h > r_\ell$. This opens up the possibility that firms mix between prices in the interval $[r_\ell, r_h]$ and prices below r_ℓ . Type ℓ consumers would then search when getting a first price quote above r_ℓ . This in turn might make it profitable for firms to price below r_ℓ to prevent such search. However, individual demands should be chosen appropriately to make sure that type r_ℓ consumers' demand is sufficiently elastic at prices below r_ℓ so firms would choose to price strictly below r_ℓ with positive probability: otherwise r_ℓ could not be a reservation price. Assuming that this can be achieved, we conjecture the equilibrium would exhibit a comparative statics pattern with respect to search costs similar to the one described in the homogeneous consumer case with multiple peaks: with firms being more likely to charge high prices if the search cost is smaller. The idea is that, if the search cost became smaller, then the probability of a low price below r_ℓ should become sufficiently small in order for r_h to remain sufficiently high as compared to r_ℓ so a firm finds it profitable not to prevent consumers with reservation price r_ℓ from searching. The example below illustrates a similar logic in a somewhat different environment.

Interestingly, Diamond himself (Diamond, 1987) provides an intriguing twist whereby demand heterogeneity can overcome the Diamond paradox. The reason this works is because he assumes search costs are delay costs instead of a cost per search (we have been supposing

the latter throughout). It could be, for instance, that consumers are hit by direct response advertising while they are surfing on the web or watching TV as in Renault (2016): the delay is then determined by the frequency of advertisements from firms selling competing products. The following simplified version of his model illustrates. Consumers have unit demands and there are only two consumer types: a measure m_ℓ of them have valuation v_ℓ and a measure m_h have valuation v_h , $0 < v_\ell < v_h$, $m_\ell + m_h = 1$. Production costs are zero. In a standard sequential search setting with cost per search $s > 0$, the paradox would sustain because the lowest equilibrium price could not be below v_ℓ , and hence low valuation consumers would never search. Suppose instead that the only cost associated with searching is due to having to wait before being able to consume the product, so the corresponding surplus is discounted by a factor $\delta \in [0, 1]$. Then, a low valuation consumer will not buy if she runs into a price strictly above v_ℓ , but she will buy if she later runs into a product priced at v_ℓ (we assume that she buys if indifferent) and waiting involves no cost because her surplus was zero to begin with. We seek a symmetric two-price equilibrium, with firms mixing between v_ℓ and $p_h > v_\ell$ and the probability of pricing low is α . The higher price cannot exceed v_h . Furthermore, in order to generate some positive profit, it must be low enough to stop a high valuation consumer from searching on in hope of getting the lower price (here we assume that if a consumer finds a high price after searching she buys from the second firm searched). Hence, it must be less than some reservation price r at which a consumer is just indifferent between searching and not searching. This reservation price is solution to

$$(1 - \delta)(v_h - r) = \delta\alpha(r - v_\ell).$$

The left-hand side represents the cost of searching associated with having to consume the product later if the consumer ended up not getting a lower price. The right-hand side is the expected benefit of search, measured by the discounted expected decrease in price. Thus $r = \frac{v_h - \delta(v_h - \alpha v_\ell)}{1 - \delta(1 - \alpha)}$. Because r is less than v_h and demand is perfectly inelastic it is optimal for a firm choosing the higher price to charge r (where again a consumer buys if indifferent). In particular, because high valuation consumers do not search in equilibrium, undercutting r would not be an optimal deviation. A high valuation consumer buys from a firm if and only if she starts out getting a quote from that firm, whereas a low valuation consumer buys from a firm if that firm charges v_ℓ and, if she starts off at that firm or the other firm is charging r . Search is random, so each firm initially gets half of the consumers.

Appendix C establishes that, if parameters are such that v_h is the monopoly price, as long as there is not too much discounting (i.e., δ is large enough), then there exists a unique $\alpha^* \in [0, 1]$ that characterizes a mixed strategy equilibrium in prices. There are also two pure strategy equilibria: one where both firms charge the monopoly price v_h and one where they both charge v_ℓ . If the delay between the two periods is reduced so δ increases, r decreases, whereas, for a given probability α , profit associated with the low price increases, because the value of selling to additional consumers in the second period is increased. As a result, the equilibrium involves a lower probability α^* that firms charge a low price. This setting thus provides a simple theoretical underpinning for the empirical findings by Ellison and Fisher Ellison (2014) that the price of rare used books sold online is higher than it is for the same items sold offline: interested buyers enjoy searching for such products so they are all shoppers but they enjoy getting hold of them as early as possible and, in an equilibrium where high prices and low

prices coexist, the probability of finding a low price must be low in order for high valuation buyers to be willing to pay a high price while they expect to get frequent price quotes.⁵¹

We next turn to search cost heterogeneity.

5.2 Search Cost Heterogeneity

In this subsection we concentrate on settings with a homogeneous product. Salop (1977) shows that a “noisy” monopolist can effectuate price discrimination by selling the same good at multiple outlets and, by setting different prices at each, discriminate against those with high search costs. As he says (p. 393) “dispersion acts as a costly device for sorting consumers into sub-markets to permit price discrimination.” An obvious first question is whether search cost heterogeneity can overcome the Diamond paradox for single-product firms. As we now argue, the answer is mixed.⁵²

A first crucial remark is that the Diamond paradox does not require that all consumers have the same search costs. All the arguments go through if there is a strictly positive lower bound on search costs. In order to obtain pricing below the monopoly price (and possibly some price dispersion), it is necessary that the search cost distribution has a fat enough tail in the neighborhood of zero. Some analysis of search market equilibrium with general search cost distributions can be found in Rob (1985) for monopolistic competition, and Stahl (1996) for oligopoly. It is quite challenging to get substantive results, especially for oligopoly. Still, a few takeaways from the analysis in Stahl (1996) are worth noting. First, it is possible to have a one-price equilibrium at a price below the monopoly price. This is because, if the density of search costs at zero is positive, and if all firms are expected to charge some price p^* , then a firm’s demand at p^* is very elastic for upward changes in price because all consumers with search costs near zero would want to switch to other sellers if the firm deviates upwards. Still the monopoly pricing equilibrium always exists as long as there is no atom at zero search cost. This equilibrium is unique if the density at zero is zero. Importantly, if the search cost distribution is atomless and its hazard rate increases over its support, then any symmetric Nash equilibrium is in pure strategies, implying that no consumer searches beyond the first firm encountered, expecting the same price at all firms. Hence (at least if we restrict attention to symmetric equilibria), there can be search in equilibrium only with some mass points in the search cost distribution (in particular at zero) or possibly if the distribution has a non-increasing hazard rate. Little is known about the properties of such equilibria with dispersed prices, except in the rather special case we discuss below in more detail.

A simple way to avoid a strictly positive lower bound on search costs is to assume as in subsection 2.1 that there is an atom of shoppers with zero search costs and the rest of the consumer population shares the same search cost $s > 0$. This simple modification of the basic framework of Varian (1980) is analyzed by Stahl (1989). The model is solved assuming that shoppers search through all the firms even if they expect the same price everywhere, and

⁵¹ The empirical analysis in Ellison and Fisher (2014) is structural and relies on a very different model, which is closer to that of Baye and Morgan (2001).

⁵² An early example is the tourists-and-natives model of Salop and Stiglitz (1977), which is closed with free entry of firms facing U-shaped average costs. A fraction of firms are bargains, pricing at minimum average cost, and found by lucky tourists and all natives (shoppers). The rest are rip-offs, pricing at the reservation price, and serving those unlucky tourists who happen upon them first, and for whom search costs are too high to countenance seeking a bargain.

non-shoppers can observe a first price for free. This atom of shoppers implies an atomless equilibrium distribution of prices if we restrict attention to equilibria where all firms play the same strategy.⁵³ This is because, if there were an atom at \hat{p} in the price distribution, then a firm charging \hat{p} could lower its price slightly, thus increasing the probability that it captures the entire shopper population by a strictly positive amount and therefore earn a strictly larger profit than at \hat{p} , which contradicts the assumption that \hat{p} is played with a positive probability in equilibrium.

Stahl's original paper allowed for downward-sloping individual demands. However, if we assume unit demands for consumers, we can engage the results of the Varian (1980) model presented in subsection 2.1. In that model, there is an exogenous reservation price, r , and “captive” consumers are exogenously allocated symmetrically to firms. We can now take that framework and endogenously determine r as the price at which a consumer with a search cost, s , will be indifferent as to searching again, given the expected prices of firms; and we break indifference so that the consumer does *not* search. Then, the “captive” consumers of the Varian model are analogous to the costly search types who randomly select a particular firm to start at. Because the price distribution is atomless, if the highest price was strictly above r , a firm charging that price could not sell and so we know that, as in subsection 2.1 r is the largest price. Then, as explained in subsection 3.1 the reservation price is the sum of search cost and expected price.

Drawing on Janssen et al. (2005), we show in Appendix D that the equilibrium reservation price (for the model with unit consumer demand) is uniquely determined and is proportional to s , with the equilibrium distribution of prices given by (8.1) in subsection 2.1.⁵⁴ Janssen et al. (2005) show that the equilibrium reservation price here unambiguously rises with n , which reflects the property that the price per firm rises with n in the sense of first-order stochastic dominance. We also show in Appendix D the striking result that the expected minimum price is independent of n (Janssen et al., 2011). Even though consumers face higher prices at each individual firm, the greater number of options exactly cancels this out. The implications of more competition are therefore different for different consumer groups (Stahl, 1989, also pointed out tensions in his setting with price-sensitive demand). The shoppers are unaffected, but the others are strictly worse off (because the first search is expected to find a higher price to start with, and no search ensues). These results are usefully compared to those for Varian's (1980) model, in which the reservation price is fixed exogenously. As we pointed out earlier, the shoppers are better off there because the minimum price falls (while the others are worse off, as here). The difference in results comes from the endogenous reservation price, which rises with n in the search context and so brings up the minimum price.

This model has become a very useful workhorse for analyzing search equilibrium in markets for homogeneous products. In particular, it does capture the increased competitiveness in a market resulting from a lower search cost because r is decreasing in s so a decrease in search costs induces a new price distribution that is stochastically dominated in the first order by the old one. Although this decrease in prices arises because search becomes more attractive, it does not result from consumers having better information about the competing prices: in

⁵³ Varian's (1980) model has multiple asymmetric equilibria too, with any number of two or more firms playing a mixed strategy, and the rest playing r (Baye, Kovenock, and de Vries, 1992, Kocas and Kiyak, 2006). This feature carries through to the Stahl-type analysis with endogenous r .

⁵⁴ Specifically, they show that $r = \frac{s}{1-\alpha}$, where $\alpha = \int_0^1 \frac{1}{\frac{\sigma}{m-\sigma}ny^{n-1}+1} dy$.

equilibrium, non-shoppers know only one price and shoppers know all prices independent of the level of the search cost. Thus, this setting fails to deliver usable predictions about the evolution of consumer search behavior and hence of their information from improved information technologies.⁵⁵

We next consider heterogeneity of search costs when consumers search for a good match.

5.3 Intensive and Extensive Margins

We now allow for search cost heterogeneity to see how equilibrium prices react to lower search costs. From the analysis in section 3, we expect consumers to search more when search costs less, inducing more competition and hence lower prices. This, however, overlooks the impact of lower search costs on a consumer's decision to search actively. If the search costs go down for consumers with high search costs to induce them to become active in new markets, the search cost distribution of participating consumers in these markets can rise to induce prices to rise. Janssen et al. (2005) illustrate this point within the framework of Stahl (1989), described in subsection 5.2. They show that if the search cost of non-shoppers is large enough, they are indifferent between participating in the market or not. Their equilibrium participation probability falls as their search costs decreases,⁵⁶ so that the share of shoppers in the market decreases, resulting in higher prices. We now discuss, following Moraga-González, Sandor, and Wildenbeest (2017), how this idea can be fruitfully generalized within a search framework with horizontal differentiation à la Wolinsky (1986).

Consider again the monopolistic competition setting of subsection 3.2, which is now extended to include a continuum of consumers with search costs distributed on $[\underline{s}, \bar{s}]$, $\underline{s} > 0$. Heterogeneity in search costs impacts the firm's demand in a non-trivial way. For each search cost level, s , there is a different threshold match $\hat{x}(s)$ that determines the consumer's demand. Thus a firm's demand integrates the probability of a purchase over all values of $\hat{x}(s)$ so the standard existence argument based on an increasing hazard rate for the match distribution does not work here. However, as Moraga-González et al. show, this consumer heterogeneity can be dealt with using a fairly standard argument. The intuition for this is as follows. Recall from subsection 3.2 that under monopolistic competition, each firm only competes with the option of searching on, as represented by the threshold $\hat{x}(s)$. This is just as if each firm faced a single competitor for whose product the consumer's match is $\hat{x}(s)$. Hence the arguments from Caplin and Nalebuff (1991) showing equilibrium existence in an oligopoly model where matches have log-concave densities can be applied here to show existence as long as $\hat{x}(s)$ has a log-concave density.⁵⁷

To build some intuition about the comparative statics regarding changes in the search cost distribution, consider a simple example with only three search cost levels, which are initially $s_{\ell 1}$, s_{m1} and s_{h1} and change to $s_{\ell 2}$, s_{m2} and s_{h2} . The fraction of consumers at each level are first $\pi_{\ell 1}$, π_{m1} , π_{h1} , and then $\pi_{\ell 2}$, π_{m2} , and π_{h2} , for low, intermediate, and high search costs

⁵⁵ Bénabou (1993) argues that this unappealing feature can be overcome by introducing marginal cost heterogeneity as in Reinganum (1979). Low search cost consumers keep on searching until they find a firm with a low enough cost. However, as in Reinganum's setting, equilibrium prices for low-cost firms are monopoly prices and bringing the search cost distribution down to zero cannot bring prices below the monopoly price of the most efficient firm.

⁵⁶ This property is rather a vagary of the mixed strategy participation indifference condition.

⁵⁷ Moraga-González et al. do not, however, provide general joint conditions on the search cost distribution and the match distribution guaranteeing that $\hat{x}(s)$ has a log-concave density.

respectively. The new search costs are stochastically lower than the original ones (in terms of first-order stochastic dominance). Furthermore, assume that only consumers with search costs of at most s_{m1} engage in search initially. Because consumers of types ℓ and m have lower search costs after the change, they search more and this per se induces downward pressure on prices. This is what Moraga-González et al. call the *intensive margin*. If type h consumers remain out of the market after the drop in search costs, then only the intensive margin is in play, and the price goes down. If, however, search costs for some type h consumers falls by enough that they search at the new equilibrium, then the composition of the participating consumers changes in a way that might upset the stochastic decrease in search costs for active searchers. This change in the *extensive margin* might pull prices upwards.

Which margin (extensive or intensive) dominates depends on how the search cost distribution changes. To illustrate, suppose the only change is a drop of the high search cost (as in the Janssen et al., 2005 example discussed above), so that s_{h2} is now low enough to induce search by the high types. Then search costs for the participating consumers unambiguously increase in a stochastic sense, so that the extensive margin dominates, leading to an increase in price. Alternatively, suppose instead that search values do not change but consumer fractions do. If some of the high types shift to the middle type with no other change (so $\pi_{m2} > \pi_{m1}$ and $\pi_{\ell2} = \pi_{\ell1}$) then the new distribution for participating consumers is stochastically higher. Once again, the impact of the extensive margin dominates. In order for the intensive margin to be the dominant factor it is necessary that there is a sufficient shift of intermediate search cost consumers to low search cost so that $\frac{\pi_{\ell2}}{\pi_{m2}} > \frac{\pi_{\ell1}}{\pi_{m1}}$ (which ensures that the proportion of low search cost consumers among searching consumers has increased). Thus, the intensive margin will dominate if the shift of consumers towards the lower values of search costs dominates the shift in favor of higher search cost values. This can be captured by comparing the two distributions in terms of a monotone likelihood ratio property. Moraga-González et al. show that, if the likelihood ratio between the final distribution and the initial distribution⁵⁸ is increasing, then the stochastic drop in search costs will induce a price increase, whereas if it is decreasing, then lower search costs will result in a price fall.⁵⁹

As noted by Moraga-González et al., the study by Hortaçsu and Syverson (2004) on the US mutual fund market in the late 1990s provides an empirical illustration of how the extensive margin may lead to higher prices in a market where search costs fall due to the development of the Internet. They find that prices increased and, according to their estimates, search costs dropped at the lower percentile of the distribution but actually rose at the upper percentiles. Hortaçsu and Syverson explain this by the arrival of new high search cost households in the mutual fund market, whose search costs went down so much that they found it worth looking for investment opportunities.

5.4 Ordered Search

We treat here ordered search when buyers have heterogeneous search costs. Let all consumers have unit demands for a homogeneous good with common valuation v , assumed to be large. Firms are searched in an exogenous order, the same for all consumers (think geography, for example, with all consumers starting at the same firm on a street, and getting further away

⁵⁸ This is defined as the ratio of the two densities.

⁵⁹ Their condition implies this in the case of a stochastic decrease in search costs.

from the start point with successive firms searched). What we shall show is that consumers with lower search costs will search longer (i.e., further). Notice before we start that the only way consumers will want to keep searching is if they expect lower prices later. This means that, as per the geography example, consumers cannot choose a later firm without going past an earlier one – if instead a consumer could reach any firm with an identical cost of search, she would want to skip immediately to the end.⁶⁰ The analysis that follows is based on Arbatskaya (2007): we specialize the search cost distribution to a uniform one, with $s \in [\underline{s}, \bar{s}]$, and we set $\bar{s} = \underline{s} + 1$.

To illustrate, suppose there are three firms with zero production costs and they are searched in order 1 through 3. Consumers who observe p_2 at firm 2 will go through to firm 3 if they expect a price discount there greater than their search cost. That is, letting p_i^* be firm i 's equilibrium price, they continue if $p_2 - p_3^* > s$. Now consider a consumer observing p_1 at firm 1. She will search on to firm 2 while expecting to buy from it at the equilibrium price p_2^* if $s < p_1 - p_2^*$. Hence, in order for firm 2 to be active in equilibrium, prices must satisfy $p_1^* - p_2^* > p_2^* - p_3^*$, or, any consumer who would move on to firm 2 expecting to buy at firm 2 would end up preferring to buy from firm 3. The argument easily generalizes to n firms and Arbatskaya (2007) shows that price differences decrease for all active firms (which must be the earliest firms in the sequence). Thus it is necessarily the consumers with the highest search costs that stop and buy at a given firm. It also means that, in equilibrium, the search decision is myopic in the sense that a consumer moves on to the next firm if and only if it would be optimal to do so even if the next firm was the last. In general, if consumers do not search in the optimal order characterized by Weitzman (1979) as is the case here, optimal search is not necessarily myopic. Myopia here is an equilibrium property in the sense that it would not be optimal to search myopically if price differences were not decreasing.

Further note that firm 3 cannot be active in equilibrium. If it expected any visit, it would just undercut firm 2's price and could thus capture all visiting consumers: then it would not be optimal for any consumer to search on all the way to firm 3 because no price discount can be expected there. Consider now firm 2's problem assuming that all consumers with search costs in excess of $\underline{s} + \Delta$ with $\Delta > 0$ stop at firm 1. Firm 2 cannot affect this decision to stop at firm 1, which depends on its expected price rather than on its actual price.⁶¹ Because $s \geq \underline{s}$, firm 2's profit is $p_2 (\underline{s} + \Delta - (p_2 - p_3^*))$ for $(p_2 - p_3^*) \geq \underline{s}$ and $p_2 \Delta$ otherwise. The solution must be to set $p_2^* = p_3^* + \underline{s}$ so that further search is deterred. Hence the price elasticity of demand at that price for a price increase must be at most -1 , that is, $p_3^* + \underline{s} \geq \Delta$. The lowest expected equilibrium price for firm 3 such that firm 2 chooses to deter further search is $p_3^* = \Delta - \underline{s}$, and the associated profit-maximizing price for firm 2 is $p_2^* = \Delta$. Any $p_3^* \geq 1 - \underline{s}$ constitutes an equilibrium, with $p_2 = p_3^* + \underline{s}$ so that there is a continuum of equilibria. Following Arbatskaya (2007), we choose the lowest possible value for p_3^* consistent with no search at the final firm.⁶² The last firm, though inactive, performs a policing role on its predecessors.

⁶⁰ Notice that the Stahl model with ordered search has the same property: it can be readily shown for duopoly that the second firm has a lower equilibrium expected price than the first one: if it were possible for consumers to choose which firm to sample first, they would choose the second one, upsetting the equilibrium pricing.

⁶¹ Arbatskaya (2007) also analyzes the case when all prices are observed (so that a firm influences both margins). There is some literature on observed price deviations, notably Carlson and McAfee (1983).

⁶² Alternatively, its price might be determined by local demand conditions. Or, indeed, one might also include some shoppers, and the firms far back would only encounter them, and so set price equal to marginal cost.

Because we seek an equilibrium where firm 2 is active, firm 1 must be setting a price above the kink in its demand corresponding to the price at which all consumer would give up search. This means that it selects a price that yields a unit price elasticity. Its demand being $1 - \Delta$, its equilibrium price is $p_1^* = 1 - \Delta$. Hence the search cost such that a consumer is indifferent between buying from firm 1 and searching is $p_1^* - p_2^* = 1 - 2\Delta$ and, because firm 2 sells to all consumers with search costs below that threshold, we have $\Delta = 1 - 2\Delta - \underline{s}$, which yields

$$\Delta = \frac{1 - \underline{s}}{3}. \quad (8.12)$$

It follows that firm 2 can be active only if $\underline{s} < 1$. Note further that if \underline{s} is more than $\frac{1}{4}$, $p_3^* = \Delta - \underline{s} < 0$, which is optimal because no consumer visits firm 3, although it is a weakly dominated strategy. Interestingly, a drop in the minimum search cost has diverging impacts on the two prices: it increases the price of the second firm and decreases that of the first firm.

The insight that the maximum number of active firms in equilibrium decreases if the minimum search cost increases extends to the n firm case. An equilibrium where $n - 1$ out of n firms are active is characterized by $n - 1$ market shares $\Delta_1, \dots, \Delta_{n-1}$ with $\Delta_i > 0$ for all i . Because pricing by firms 1 through $n - 1$ must set the price elasticity of demand to -1 , we have $p_i^* = \Delta_i$. These market shares constitute an equilibrium if and only if they satisfy $\sum_{i=1}^{n-1} \Delta_i = 1$ and $\Delta_i = \underline{s} + \sum_{j=i+2}^{n-1} \Delta_j + 2\Delta_{i+1}$ for $i = 1, \dots, n - 2$. The second equality merely means that consumers who buy from firm i are those who do not buy from earlier firms and have a search cost larger than $p_i^* - p_{i+1}^* = \Delta_i - \Delta_{i+1}$. It also guarantees the equilibrium property that price differences should decrease and should be equal to some threshold search cost lying above \underline{s} . It is readily seen that the larger is n the lower \underline{s} should be in order for the two constraints to hold simultaneously. Furthermore, for $\underline{s} = 0$, it is always possible to specify a sequence of market shares satisfying the two constraints for any number of active firms: the second constraint allows for deriving all market shares from the market share of the last active firm Δ_{n-1} so that taking it appropriately small ensures that the sum of market shares adds up to 1.

What is in evidence here for $\underline{s} > 0$ is a “natural oligopoly” (or finiteness) property of the ordered search model. Natural oligopolies were earlier described by Shaked and Sutton (1983) for vertical differentiation models. Even as fixed costs tend to zero, only a finite number of firms might survive in equilibrium because lower qualities cannot turn a profit due to their disadvantage. So it is here that firms too far down the search order cannot survive profitably given the pricing behavior of their predecessors.⁶³

We now return to the ranking of sellers with heterogeneous purchase probabilities and illustrate how it can be interacted with search cost heterogeneity. As discussed in subsection 4.4, Athey and Ellison (2011) consider a setting where all firms charge the same exogenous price and differ in the probability that their product matches a consumer’s need. They study how the search order can emerge as the outcome of an auction where firms bid for positions on a website. Purchase probability is private information to the product’s seller. As was pointed out in subsection 4.4, this assumption implies that higher probability firms have a higher incremental value for being ahead in the search order. Consumers should indeed expect them

⁶³ The result for $\underline{s} = 0$ parallels what obtains in the vertical differentiation setting for a standard uniform distribution of valuations for quality.

to bid more and hence, to be placed higher on the web page: so the proposed order is the optimal search order. Athey and Ellison introduce heterogeneity in search costs, which also induces a positive relation between purchase probability and seller willingness to pay for being positioned early. To see this, consider again the two firm example with an exogenous price where firms know each other's probability and assume now that there is only a fraction $\alpha \in [0, 1]$ of consumers whose search cost is low enough so that they want to check out the other product if they are not satisfied with the first product sampled. Then firm 1's incremental value for being ahead of firm 2 is $(1 - \alpha)\beta_1 + \alpha\beta_1\beta_2$, which is indeed larger than the incremental value of firm 2 for being in front of firm 1 if $\beta_1 > \beta_2$.

In addition, and maybe more importantly, heterogeneity in search costs also yields a higher total industry profit if firms are ordered according to a decreasing probability ranking, whereas the search order would not impact total profit if it were the case that all consumers searched on until finding a satisfactory product. Total profit is given by $\beta_1 + \alpha(1 - \beta_1)\beta_2$ with firm 1 searched first: for $\beta_1 > \beta_2$, this is strictly larger than with firm 2 in the top position if and only if $\alpha < 1$. Obviously, an ordering of firms with higher purchase probabilities earlier is also what is preferred by consumers, because it minimizes expected search costs (and it must arise in this fixed price setting, so consumer search behavior is consistent with the firms' bidding for prominence on the web page). The analysis of Athey and Ellison (2011) therefore yields an alignment of preferences between firms and consumers regarding the search order. By contrast, the results in Anderson and Renault (2016) where prices are endogenous, point rather to a systematic misalignment. In particular, if firms only differ by the probability of a match with the consumer's tastes, total profit maximization requires that firms with a *lower* match probability should get more prominence because early positions are priced lower. However, consumers prefer to reach high match probabilities early. Not only does this reduce total search costs, but early products are also sold at lower prices.

6 CONCLUSIONS

We began this chapter by noting that research on search costs and firm pricing has really taken off with the advent of the Internet. Search costs are very visible in searching for goods and services online. And yet, search costs were prevalent before the Internet. It might be argued that search costs were responsible for rather limited access to variety before the Internet age, and a concurrent stifling of price competition. These costs were hitherto quite hidden. Perhaps ironically, it is only when search costs have fallen that economists (Nobel Prize winners aside!) became much more aware of them (one exception is the analysis of job search, although the equilibrium analysis initiated by Diamond, 1982, only picked up in the 1990s). Along with the opening of markets has come a supply response that has greatly increased the variety of goods and services that are readily accessed.⁶⁴

We have surveyed various models of market pricing with search costs. We have noted the great achievements (much credit should be ascribed to Weitzman, 1979, for these) in the description of individual search behavior. But closing the loop and endogenizing firm pricing and realistically describing market equilibrium has proved a bit more cumbersome, though progress continues apace. Models differ by the degree of heterogeneity they assume

⁶⁴ See Goldman et al. (2010) for empirical evidence on how travel agencies and bookstores have adapted.

about individuals, and firms. For simplicity below, we divide the main approaches into the product differentiation approach following Wolinsky (1986) and the heterogeneous search costs approach involving mixed strategies for prices, following Stahl (1989) and building on Varian (1980).

Models can be judged according to various criteria. One is their success in matching empirical regularities. These include price distributions, and consumer search behavior. The Stahl model succeeds in generating a price distribution (which is a truncated Pareto for the inelastic demand case), although critics of mixed price strategy equilibrium are quite common. However, the model does not generate search in equilibrium.⁶⁵ The Wolinsky model, by contrast, involves pure price strategies (arguably a strong point per se). Under fully random search, price dispersion would have to come from asymmetries in firms' costs or qualities (which are a bit cumbersome to manage in the model), but ordered search does deliver dispersion. On the consumer side, different consumers have different search patterns, with lucky ones finding acceptable matches earlier, while others go to the end and return.⁶⁶

We can also look to the models to pull together some results about the impact of reduced search costs. In the Wolinsky (1986) model, a lower search cost raises the search threshold, so consumers become more choosy. They search longer, on average. The *total* search cost falls though under monopolistic competition for log-concave match distributions. In this case market prices fall as consumers search longer. When the number of active consumers is fixed, the price drop drives profits to fall too. This reduces total product variety in the long run so globalization can streamline (and bankrupt) local market offerings.⁶⁷

However, there is another important channel through which lower search costs might raise total variety. This is the two-sided market effect that we just suppressed by taking the number of consumers as fixed. Because lower search costs increase expected surplus from market participation, more consumers will want to look, so delivering a virtuous circle (i.e., bilateral positive externalities from participating) that induces more firms to want to partake too; thus, lower search costs generate a thickening of the market all around.⁶⁸

The two-sided market effects just discussed are one example of the externalities inherent in search markets. For another example, in the Stahl model there are externalities between the consumer groups. All consumers are better off the more informed consumers there are, as price competition for the informed is fiercer. But, the more consumers there are whose search is costly, the worse off everyone is. Anderson and Renault (2000) consider a variant of the Wolinsky model in which some consumers know their match values already, while others need to search to find them. The former immediately check out the firm that they like most, which means that their demand is more inelastic than for the other consumers. The more of

⁶⁵ The model starkly depicts two groups of consumers. One group does not search at all in equilibrium, while the other already knows all prices. Semantically, it could be argued that the latter group does (without cost) search through all products.

⁶⁶ This is less likely the bigger the number of options. All comeback consumers reach the end: extra heterogeneity of firms and/or consumers would be needed to get some consumers to return to an earlier option before reaching the end (as can be envisaged from the Weitzman, 1979, framework).

⁶⁷ Some of these contrasting effects are also at play in the (suitably extended) Stahl (1989) model. A lower search cost (for those with positive search costs) disciplines firms more and intensifies price competition (even though the actual search is unchanged, it is the threat of search that drives prices down), although lower costs per search do not impact total search costs because there is no search in equilibrium. The lower price level improves the welfare for both consumer types. Firms though suffer lower profits.

⁶⁸ This can hold too in an extended version of the Stahl (1989) model with endogenous consumer participation. With lower search costs, profits could rise in the short run and more firms enter in the long run.

them there are, the higher are equilibrium prices. These informed types impose a negative externality on all (and the uniformed imbue a positive externality). Introducing a private cost to getting informed (which differs across individuals), there is thus a negative externality from investing in information acquisition, and so *too much* information is gathered on search for product matches.

Internet search is inherently sequential, and frequently ordered to boot. Models of equilibrium pricing for such markets, with concurrent competition for positions in the search order, are only just being developed. Relatedly, the Internet also greatly facilitates access to price information (probably more so than to information about product attributes). Consumers can thus easily compare prices and firms can use pricing to direct search, with consumers being enticed to start search with the cheapest. How this affects price competition is an exciting and challenging research question: Armstrong (2016) discusses some of the recent contributions on this topic.⁶⁹ As Internet commerce develops into more different variations and opportunities, the economics of search will develop along with it. It promises to be an exciting research trajectory.

REFERENCES

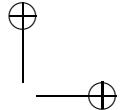
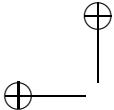
- Anderson, S.P., Renault, R. (1999). Pricing, Product Diversity, and Search Costs: A Bertrand-Chamberlin-Diamond Model. *The RAND Journal of Economics*, 30(4), 719–735.
- Anderson, S.P., Renault, R. (2000). Consumer Information and Firm Pricing: Negative Externalities from Improved Information. *International Economic Review*, 41(3), 721–742.
- Anderson, S.P., Renault, R. (2003). Efficiency and Surplus Bounds in Cournot Competition. *Journal of Economic Theory*, 113(2), 253–264.
- Anderson, S.P., Renault, R. (2016). Search Direction. Mimeo.
- Anderson, S.P., Baik, A., Larson, N. (2016) Price Discrimination in the Information Age: List Prices, Poaching, and Retention with Personalized Discounts. Mimeo.
- Anderson, S.P., De Palma, A., Nesterov, Y. (1995). Oligopolistic Competition and the Optimal Provision of Products. *Econometrica*, 63(6), 1281–1301.
- Anderson, S.-P., De Palma, A., Thisse, J.-F. (1992). Discrete Choice Theory of Product Differentiation. Cambridge, MA: MIT Press.
- Arbatskaya, M. (2007). Ordered Search. *The RAND Journal of Economics*, 38(1), 119–126.
- Armstrong, M. (2016). Ordered Consumer Search. *CEPR Discussion Paper* 11566.
- Armstrong, M., Zhou, J. (2011). Paying for Prominence. *Economic Journal*, 121(556), F368–F395.
- Armstrong, M., Vickers, J., Zhou, J. (2009a). Prominence and Consumer Search. *The RAND Journal of Economics*, 40(2), 209–233.
- Armstrong, M., Vickers, J., Zhou, J. (2009b). Consumer Protection and the Incentive to Become Informed. *Journal of the European Economic Association*, 7(2–3), 399–410.
- Athey, S., Ellison, G. (2011). Position Auctions with Consumer Search. *Quarterly Journal of Economics*, 126(3), 1213–1270.
- Bar-Isaac, H., Caruana, G., Cunat, V. (2012). Search, Design, and Market Structure. *American Economic Review*, 102(2), 1140–1160.
- Baye, M., Morgan, J. (2001). Information Gatekeepers on the Internet and the Competitiveness of Homogeneous Product Markets. *American Economic Review*, 91(3), 454–474.
- Baye, M., Kovenock, D., De Vries, C.G. (1992). It Takes Two to Tango: Equilibria in a Model of Sales. *Games and Economic Behavior*, 4(4), 493–510.
- Baye, M., Morgan, J., Scholten, P. (2006). Information, Search, and Price Dispersion. In T. Hendershott (ed.), *Handbook of Economics and Information Systems*, Vol. 1, Amsterdam: Elsevier.

⁶⁹ Price posting is also a key ingredient of the directed search models initiated by Peters (1984) and widely used in the analysis of the labor market (although Peters' original model was applied to a product market). The other key ingredient is that firms have a limited capacity, in contrast to the settings we have discussed where capacity is unlimited and marginal production costs are constant.

- Bénabou, R. (1993). Search Market Equilibrium, Bilateral Heterogeneity, and Repeat Purchases. *Journal of Economic Theory*, 60(1), 140–158.
- Bénabou, R., Gertner, R. (1993). Search with Learning from Prices: Does Increased Inflationary Uncertainty Lead to Higher Markups? *Review of Economic Studies*, 60(1), 69–93.
- Burdett, K., Judd, K.L. (1983). Equilibrium Price Dispersion. *Econometrica*, 51(4), 955–969.
- Butters, G. (1977). Equilibrium Distributions of Sales and Advertising Prices. *Review of Economic Studies*, 44(3), 465–491.
- Caillaud, B., Jullien, B. (2003). Chicken & Egg: Competition Among Intermediation Service Providers. *The RAND Journal of Economics*, 34(2), 309–328.
- Caplin, A., Nalebuff, B. (1991). Aggregation and Imperfect Competition: On the Existence of Equilibrium. *Econometrica*, 59(1), 25–59.
- Carlson, J., McAfee, R. (1983). Discrete Equilibrium Price Dispersion. *Journal of Political Economy*, 91(3), 480–493.
- Chen, Y. He, C. (2011). Paid placement: Advertising and search on the Internet. *Economic Journal*, 121, F309–F328.
- Dana, J. (1994). Learning in an Equilibrium Search Model. *International Economic Review*, 35(3), 745–771.
- De Cornière, A. (2016). Search Advertising. Mimeo, Oxford University.
- Diamond, P.A. (1971). A Model of Price Adjustment. *Journal of Economic Theory*, 3(2), 156–168.
- Diamond, P.A. (1982). Wage Determination and Efficiency in Search Equilibrium. *Review of Economic Studies*, 49, 217–227.
- Diamond, P.A. (1987). Consumer Differences and Prices in a Search Model. *Quarterly Journal of Economics*, 102(2), 429–436.
- Ellison, G., Fisher Ellison, S. (2014). Match Quality, Search, and the Internet Market for Used Books. Mimeo.
- Fershtman, C., Fishman, A. (1992). Price Cycles and Booms: Dynamic Search Equilibrium. *The American Economic Review*, 82(5), 1221–1233.
- Fishman, A. (1996). Search with Learning and Price Adjustment Dynamics. *Quarterly Journal of Economics*, 111(1), 253–268.
- Fudenberg, D., Tirole, J. (1991). Perfect Bayesian Equilibrium and Sequential Equilibrium. *Journal of Economic Theory*, 53(2) 236–260.
- Goldmanis, M., Hortaçsu, A., Syverson, C., Emre, Ö. (2010). E-commerce and the Market Structure of Retail Industries. *The Economic Journal*, 120(545), 651–682.
- Haan, M., Moraga-González, J.L. (2011) Advertising for Attention in a Consumer Search Model. *The Economic Journal* 121(552), 552–579.
- Hortaçsu, A., Syverson, C. (2004). Product Differentiation, Search Costs and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds. *Quarterly Journal of Economics*, 119(2), 403–456.
- Janssen, M., Moraga-González, J.L. (2004) Strategic pricing, Consumer Search and the Number of Firms. *Review of Economic Studies*, 71(4), 1089–1118.
- Janssen, M., Moraga-González, J.L. (2008). On Mergers in Consumer Search Markets. Mimeo.
- Janssen, M., Non, M. (2008). Advertising and Consumer Search in a Duopoly Model. *International Journal of Industrial Organization*, 26(1), 354–371.
- Janssen, M., Shelegia, S. (2016). Beliefs and Consumer Search. Mimeo.
- Janssen, M., Moraga-González, J.L., Wildenbeest, M.R. (2005). Truly Costly Sequential Search and Oligopolistic Pricing. *International Journal of Industrial Organization*, 23(5–6), 451–466.
- Janssen, M., Parakhonyak, A., Parakhonyak, A. (2016). Non-reservation Price Equilibria and Consumer Search. Mimeo.
- Janssen, M., Pichler, P., Weidenholzer, S. (2011). Sequential Consumer Search with Incompletely Informed Consumers. *The RAND Journal of Economics*, 42(3), 444–470.
- Johnson, J.P., Myatt, D.P. (2006). On the Simple Economics of Advertising, Marketing, and Product Design. *American Economic Review*, 96(3), 756–784.
- Kocas, C., Kiyak, T. (2006). Theory and Evidence on Pricing by Asymmetric Oligopolies. *International Journal of Industrial Organization*, 24(1), 83–105.
- Kohn, M., Shavell, S. (1974). The Theory of Search. *Journal of Economic Theory*, 9, 93–123.
- Lach, S. (2002). Existence and Persistence of Price Dispersion: An Empirical Analysis. *Review of Economics and Statistics*, 84(3), 433–444.
- Larson, N. (2013). Niche Products, Generic Products, and Consumer Search. *Economic Theory*, 52(2), 793–832.
- Lewis, T.R., Sappington, D.E. (1994) Supplying Information to Facilitate Price Discrimination. *International Economic Review*, 35(2), 309–327.
- Moraga-González, J.L., Petrikaitė, V. (2013). Search Costs, Demand-side Economies, and the Incentives to Merge under Bertrand Competition. *The RAND Journal of Economics*, 44(3), 391–424.
- Moraga-González, J.L., Sandor, Z., Wildenbeest, M. (2017). Prices and Heterogeneous Search Costs. *RAND Journal of Economics*, 48(1), 125–146.
- Morgan, J., Orzen, H., and Sefton, M. (2006). An Experimental Study of Price Dispersion. *Games and Economic Behavior*, 54(1), 134–158.

216 *Handbook of game theory and industrial organization: applications*

- Perloff, J., Salop, S. (1985). Equilibrium with Product Differentiation. *Review of Economic Studies*, 52(1), 107–120.
- Peters, M. (1984). Bertrand Equilibrium with Capacity Constraints and Restricted Mobility. *Econometrica*, 52(5), 1117–1127.
- Petrikaitė, V. (2016). Collusion with Costly Consumer Search. *International Journal of Industrial Organization*, 44, 1–10.
- Renault, R. (2016). Advertising in Markets. In S. Anderson, J. Waldfogel and D. Stromberg (eds), *Handbook of Media Economics, Vol 1A*, Amsterdam: Elsevier.
- Reinganum, J. (1979). A Simple Model of Equilibrium Price Dispersion. *Journal of Political Economy*, 87(4), 851–858.
- Rhodes, A., Zhou, J. (2016). Consumer Search and Retail Market Structure. Mimeo.
- Rob, R. (1985). Equilibrium Price Distributions. *Review of Economic Studies*, 52(3), 487–504.
- Robert, J., Stahl II, D.O. (1993). Informative Price Advertising in a Sequential Search Model. *Econometrica*, 61(3), 657–686.
- Rosenthal, R. (1980). A Model in Which an Increase in the Number of Sellers Leads to a Higher Price. *Econometrica*, 48(6), 1575–1579.
- Rothschild, M. (1974). Searching for the Lowest Price When the Distribution of Prices Is Unknown. *Journal of Political Economy*, 82(4), 689–711.
- Salop, S. (1977). The Noisy Monopolist: Imperfect Information, Price Dispersion and Price Discrimination. *Review of Economic Studies*, 44(3), 393–406.
- Salop, S., Stiglitz, J. (1977). Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion. *Review of Economic Studies*, 44(3), 493–510.
- Shaked, A., Sutton, J. (1983). Natural Oligopolies. *Econometrica*, 51(5), 1469–1483.
- Song, H. (2017). Ordered Search with Asymmetric Product Design. *Journal of Economics*, 121(2), 105–132.
- Stahl II, D.O. (1989). Oligopolistic Pricing with Sequential Consumer Search. *American Economic Review*, 79(4), 700–712.
- Stahl II, D.O. (1996). Oligopolistic Pricing with Heterogeneous Consumer Search. *International Journal Industrial Organization*, 14(2), 243–268.
- Stigler, G.J. (1961). The Economics of Information. *Journal of Political Economy*, 69(3), 213–225.
- Stiglitz, J.E. (1979). Equilibrium in Product Markets with Imperfect Information. *American Economic Review*, 69(2), 339–345.
- Tappata, M. (2009). Rockets and Feathers: Understanding Asymmetric Pricing. *RAND Journal of Economics*, 40(4), 673–687.
- Tirole, J. (1988). *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.
- Varian, H.R. (1980). A Model of Sales. *American Economic Review*, 70(4), 651–659.
- Weitzman, M. (1979). Optimal Search for the Best Alternative. *Econometrica*, 47(3), 641–654.
- Weyl, E.G., Fabinger, M. (2013) Pass-through as an Economic Tool: Principles of Incidence Under Imperfect Competition. *Journal of Political Economy*, 121(3), 528–583.
- Wolinsky, A. (1984). Product Differentiation with Imperfect Information. *Review of Economic Studies*, 51(1), 53–61.
- Wolinsky, A. (1986). True Monopolistic Competition as a Result of Imperfect Information. *Quarterly Journal of Economics*, 101(3), 493–512.
- Zhou, J. (2011) Ordered Search in Differentiated Markets. *International Journal Industrial of Organization*, 29(2), 253–262.
- Zhou, J. (2014). Multiproduct Search and the Joint Search Effect. *American Economic Review*, 104(9), 2918–2939.



APPENDIX A: PRICE DISPERSION

A1 Model with Exogenous Population of Captive Consumers

Let $\pi(p)$ denote a firm's expected profit when it charges a price p and all other $n - 1$ firms follow the symmetric price distribution $F(p)$. Either all other firms price above p and the firm serves all shoppers and its captives, or else at least one sets a lower price and the firm just serves its captives. Hence

$$\pi(p) = p \left\{ (\gamma + \sigma) (1 - F(p))^{n-1} + \gamma \left(1 - (1 - F(p))^{n-1} \right) \right\}.$$

Setting this equal to the profit earned by pricing at r , $\pi(r) = \gamma r$, and rearranging yields the equilibrium price distribution as given in the text.

A2 Simultaneous Search Model

Recall that v_1 is the probability that a consumer gets one quote (so that each of the n firms has an equal chance), and v_2 is the complementary probability that she gets two (and then chooses the one with the lower price). Hence

$$\pi(p) = p \left(v_1 \frac{m}{n} + 2v_2 \frac{m}{n} (1 - F(p)) \right).$$

Standard arguments show that r is in the price support and the support has no atoms. Setting $\pi(p)$ equal to the profit earned by pricing at r , $\pi(r) = rv_1 \frac{m}{n}$, and rearranging yields the equilibrium price distribution as given in the text. As per the text, the equilibrium values of v_1 and v_2 are determined by the condition that the consumer be indifferent between getting one or two quotes.

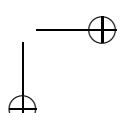
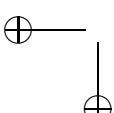
A3 Price Advertising Model

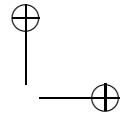
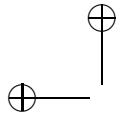
We characterize an equilibrium where a firm that does not advertise charges r and advertised prices have support $[p, r]$ where we take the convention that a consumer chooses the advertised price in case of a tie with an unadvertised one. There will be no advertising if $\frac{2A}{m} \geq r$, so assume this condition does not hold. If advertising, a firm gets profit

$$\pi(p) = pm(1 - F(p)) - A,$$

where we define $F(p)$ as the probability that a price is advertised at or below p , and so $F(r)$ is the probability of advertising. If a firm does not advertise, it gets $\frac{r}{2}m(1 - F(r))$ where $(1 - F(r))$ is the probability the other firm does not advertise (and hence the symmetric equilibrium probability). From above, $\pi(r) = rm(1 - F(r)) - A$, which we equate to the not-advertising profit to find

$$1 - F(r) = \frac{2A}{rm},$$



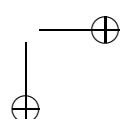
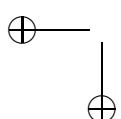


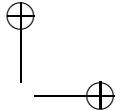
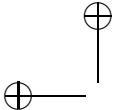
218 *Handbook of game theory and industrial organization: applications*

and the equilibrium profit is consequently A . The equilibrium probability of an advertised price at or below p is then

$$F(p) = 1 - \frac{2A}{mp}$$

for $p \in \left[\frac{2A}{m}, r \right]$.





APPENDIX B: HORIZONTAL DIFFERENTIATION AND SEARCH IN MONOPOLISTIC COMPETITION

B1 Profit Function Quasiconcavity and Symmetric Equilibrium

Given the demand function from the text, we have firm i 's profit as

$$\pi_i(p_i) = (p_i - c)m \frac{1 - F\left(\hat{x} + \frac{p_i - p^*}{\mu}\right)}{1 - F(\hat{x})},$$

which is positive for $p_i \in [c, \mu(b - \hat{x}) + p^*]$, and, since any candidate p^* exceeds c , this interval is non-empty. First note that if $c < \mu(a - \hat{x}) + p^*$ then profit is linearly increasing in p_i for $p_i \in [c, \mu(a - \hat{x}) + p^*]$ because all consumers reaching firm i stop at it ($F\left(\hat{x} + \frac{p_i - p^*}{\mu}\right) = 0$). Over the interval $p_i \in [\max\{c, \mu(a - \hat{x}) + p^*\}, \mu(b - \hat{x}) + p^*]$, we can write

$$\frac{d\pi_i(p_i)}{dp_i} = \frac{mf\left(\hat{x} + \frac{p_i - p^*}{\mu}\right)}{1 - F(\hat{x})} \left\{ \frac{1 - F\left(\hat{x} + \frac{p_i - p^*}{\mu}\right)}{f\left(\hat{x} + \frac{p_i - p^*}{\mu}\right)} - (p_i - c) \right\}.$$

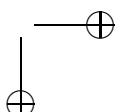
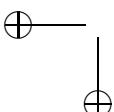
The term outside the parentheses is positive as long as demand is positive, so consider the terms inside. The first (ratio) term is positive, and decreasing in p_i from the increasing hazard rate property (log-concavity of $1 - F(\cdot)$), from which we subtract the mark-up term, $(p_i - c)$, which is zero at $p_i = c$ and linearly increasing in p_i . Hence, the profit derivative is at first positive then it is negative. Profit is therefore quasiconcave. The profit-maximizing best response to a candidate $p^* > c$ is therefore either at $\mu(a - \hat{x}) + p^*$ or else where the term in parentheses is zero. The former cannot constitute a symmetric equilibrium, so the symmetric equilibrium is where the term in parentheses is zero at $p_i = p^*$, which is the expression given in the text.

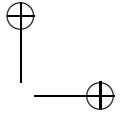
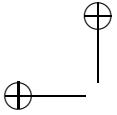
B2 Behavior of Equilibrium Price for μ Low Enough

First recall from (8.6) that $\mu \int_a^b (\epsilon_i - \hat{x}) f(\epsilon_i) d\epsilon_i = s$ defines the match-value stopping rule, \hat{x} . Then the value $\underline{\mu} = \frac{s}{\int_a^b (\epsilon_i - a) f(\epsilon_i) d\epsilon_i} > 0$ is the value of μ that just entails the consumer stopping immediately, even at the worst possible match, a . So, for $\mu < \underline{\mu}$ the consumer stops immediately. The equilibrium price (from the text) is $p^* = c + \frac{\mu[1 - F(\hat{x})]}{f(\hat{x})}$, which is finite as long as $\hat{x} > a$. However, as $\mu \downarrow \underline{\mu}$, this price tends to $c + \frac{\mu}{f(a)}$, which tends to infinity for $f(a) = 0$, which implies p^* must be decreasing over some non-zero measure set in the neighborhood of $\underline{\mu}$.

B3 Total Search Costs

We wish to determine whether or not total search costs spent increase or decrease when s falls. For example, is more time spent searching? The equilibrium chance of a consumer stopping





(and therefore buying) at any firm, conditional on reaching it, is $1 - F(\hat{x})$. The expected number of searches is then¹

$$E = \frac{1}{1 - F(\hat{x})},$$

so that the equilibrium total search cost is

$$sE = \frac{\mu \int_{\hat{x}}^1 (1 - F(x)) dx}{1 - F(\hat{x})}.$$

Assuming that $1 - F$ is log-concave, then the integral here is also log-concave (by the inheritance property of log-concave functions under integration: see e.g., Caplin and Nalebuff, 1991). Then the ratio $\frac{-(1-F(\hat{x}))}{\int_{\hat{x}}^1 (1-F(x)) dx}$ should be decreasing in \hat{x} , so that sE is decreasing in \hat{x} . Then, note from the search rule that²

$$\frac{d\hat{x}}{ds} = \frac{-1}{\mu(1 - F(\hat{x}))}$$

so that sE increases in s .

B4 Equilibrium Analysis Under Asymmetric Costs

Firm i 's demand is proportional to $1 - F(p_i + \hat{u})$, so firm i 's profit is proportional to

$$\pi_i(p_i) = (p_i - c_i)(1 - F(p_i + \hat{u})).$$

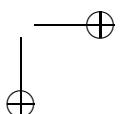
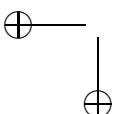
The argument in the first subsection of this Appendix applies, *mutatis mutandis*, to show that π_i is quasiconcave in p_i , so that an interior solution (i.e., with $p^* \in [a - \hat{u}, b - \hat{u}]$) solves the first-order condition

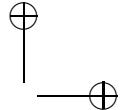
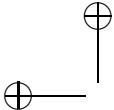
$$p^*(c_i, \hat{u}) = c_i + \frac{(1 - F(p^*(c_i, \hat{u}) + \hat{u}))}{f(p^*(c_i, \hat{u}) + \hat{u})}.$$

The increasing hazard rate property (that demand, $1 - F(\cdot)$, is strictly log-concave) implies that $k = -\left[\frac{(1-F(x))}{f(x)}\right]' > 0$. Hence, using the implicit function theorem that the cost pass-through rate, $\frac{dp^*(c, \hat{u})}{dc} = \frac{1}{1+k} \in [0, 1]$. Similarly, $\frac{dp^*(c, \hat{u})}{du} = \frac{-k}{1+k} \in [-1, 0]$, as claimed in the text.

¹ The expected number of searches is (dropping temporarily the dependence of F on \hat{x}) $E = (1 - F) + 2(1 - F)F + 3(1 - F)F^2 + \dots$ The series $1 + 2F + 3F^2 + \dots$ is the sum of the series $S = 1 + F + F^2 + \dots$ plus F times the series S plus F^2S etc., i.e., $\frac{S}{1-F}$, and likewise $S = \frac{1}{1-F}$, so that $E = \frac{1}{1-F}$.

² Notice too that the change in the expected number of searches with respect to s is $\frac{dE}{ds} = \frac{-f(\hat{x})}{(1-F(\hat{x}))^2} < 0$. The "demand" for search therefore slopes down.





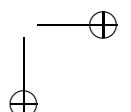
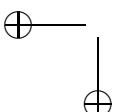
Finally, briefly consider non-interior solutions, and note the profit derivative (from the quasi-concave profit function) is

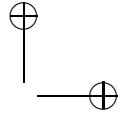
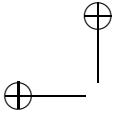
$$\frac{d\pi_i(p_i)}{dp_i} = -(p_i - c_i)f(p_i + \hat{u}) + (1 - F(p_i + \hat{u}))$$

A firm will choose to sell to all comers if this is negative where $p_i + \hat{u} = a$: rearranging, the condition for this to happen is

$$c_i \leq a - \hat{u} - \frac{1}{f(a)}.$$

On the other side, a firm will not want to sell at all if the profit derivative is positive where $p_i + \hat{u} = b$; rearranging, this holds if $c_i \geq b - \hat{u}$. In this case the firm cannot capture even its most eager customer by pricing at marginal cost.





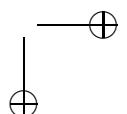
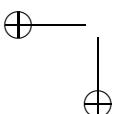
APPENDIX C: WHEN SEARCH COSTS ARE WAITING COSTS OR SEARCH WITH DISCOUNTING

Given the equilibrium strategy played by its competitor, a firm is indifferent between the two prices if

$$\frac{m_h}{2} \left(\frac{v_h - \delta(v_h - \alpha v_\ell)}{1 - \delta(1 - \alpha)} \right) = \frac{\alpha}{2} v_\ell + (1 - \alpha) ((1 + \delta)m_\ell + m_h) \frac{v_\ell}{2}. \quad (\text{C1})$$

where firms are assumed to have the same discount rate as consumers.

The left-hand side is strictly decreasing and convex in α : it is equal to $\frac{m_h v_h}{2}$ for $\alpha = 0$ and $\frac{m_h}{2}(v_h - \delta(v_h - v_\ell))$ for $\alpha = 1$. The right-hand side is linear and decreasing in α from $((1 + \delta)m_\ell + m_h) \frac{v_\ell}{2}$ for $\alpha = 0$ to $\frac{v_\ell}{2}$ for $\alpha = 1$. Because v_h is the monopoly price we have $\frac{m_h v_h}{2} > \frac{v_\ell}{2}$. Assume m_ℓ is small enough so that $\frac{m_h v_h}{2} > (m_\ell + \frac{m_h}{2}) v_\ell$, and, in addition, that δ is close enough to 1 so that $\frac{m_h}{2}(v_h - \delta(v_h - v_\ell)) < \frac{v_\ell}{2}$. Then there are two symmetric pure strategy equilibria where price is v_h and v_ℓ respectively. In addition, there exists a value of α , $\alpha^* \in [0, 1]$ that satisfies (C1). Furthermore, there can be only one solution: if there were a second solution, then the left-hand side of (C1) would cross the right-hand side from below and, since the former is convex and the latter is linear, the right-hand side profit would have to be larger than the left-hand side profit at $\alpha = 1$, which is not the case for δ large enough.



APPENDIX D: SEARCH EQUILIBRIUM WITH SOME SHOPPERS

We follow the exposition of Stahl's model (specialized to unit demand) in Janssen et al. (2005). These authors decompose the analysis into two constituent parts. On the firm side, the equilibrium relation for the price distribution is given from the Varian model (see Appendix A). The reservation price, r , is endogenously determined from the condition that the consumers with search costs stop at first search. That is, r is determined from the condition that the expected benefit from further search be just equal to the search cost, s (so that no such consumer searches again):

$$r - \int_{\underline{p}}^r pf(p) dp = s$$

with $\underline{p} = \frac{\gamma}{\gamma+\sigma} r$ (as per the Varian model). Here the left-hand side is the expected price saved from a further search, because the integral expression is the expected price at any randomly chosen firm and can be rewritten as $\int_{\underline{p}}^r pf(p) dp = \int_0^1 pdf(p) dp$.

As we showed in Appendix A, the Varian model gives $(1 - F(p)) = \left(\frac{\gamma}{\sigma} \left(\frac{r}{p} - 1\right)\right)^{\frac{1}{n-1}}$ with $\gamma = \frac{m-\sigma}{n}$ (below we break out n when pertinent in order to discuss the effects of changing the number of firms). We now show that there is a unique r solving the model, and give its closed-form solution. Indeed, set $y = 1 - F(p)$ to invert the equilibrium distribution characterization as

$$p = \frac{r}{\frac{\sigma}{\gamma} y^{n-1} + 1}$$

which we can then insert in the expected price expression to write

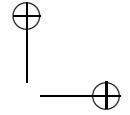
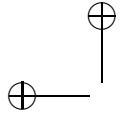
$$\int_0^1 pdf(p) dp = r \int_0^1 \frac{1}{\frac{\sigma}{\gamma} y^{n-1} + 1} dy, \quad (D1)$$

and hence we have an explicit solution for the reservation price, as given in the text, namely $r = \frac{s}{1-\alpha}$ (with $\gamma = \frac{m-\sigma}{n}$ and $\alpha = \int_0^1 \frac{1}{\frac{\sigma}{\gamma} y^{n-1} + 1} dy$).

The insight of Janssen et al. (2011) to find the expected minimum price, Ep_{\min} , in the market follows from the indifference property of the mixed strategy equilibrium. Charging the reservation price nets a firm $\frac{r(m-\sigma)}{n}$, which must equal its overall expected profit. The latter has two sources, the expected price on its share of consumers with positive search costs (i.e., $\alpha r \frac{m-\sigma}{n}$), plus its chance ($1/n$) of getting the s shoppers at the expected minimum price. Pulling that together yields the indifference condition as

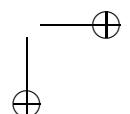
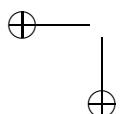
$$r(m-\sigma) = \alpha r(m-\sigma) + \sigma Ep_{\min},$$

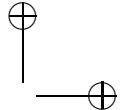
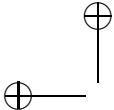
which rearranges to $Ep_{\min} = \frac{r(m-\sigma)(1-\alpha)}{\sigma} = \frac{(m-\sigma)s}{\sigma}$, where the last step recalls the reservation price property that $r = \frac{s}{1-\alpha}$. The expected minimum is clearly independent of n .



224 *Handbook of game theory and industrial organization: applications*

Finally, it is useful to draw out the simpler case for duopoly. From the analysis above, with $n = 2$, we have (by integrating) $\alpha = \frac{m-\sigma}{2\sigma} \ln \left(\frac{m+\sigma}{m-\sigma} \right)$, $r = \frac{s}{1-\alpha}$, and $(1 - F(p)) = \frac{\gamma}{\sigma} \left(\frac{r}{p} - 1 \right)$, with $p = r \frac{m+\sigma}{m-\sigma}$. The distribution equation is already in the Varian model analysis, so let us simply derive the expected price from a search, which is the left-hand side of (D1). Noting from the Varian distribution result that $f(p) = \frac{\gamma}{\sigma} \frac{r}{p^2}$, we have $\int_0^1 p dF(p) = \int_{\underline{p}}^r p f(p) dp = r \frac{m-\sigma}{2\sigma} \ln \left(\frac{m+\sigma}{m-\sigma} \right)$, and hence the results above are readily verified.





9. Market structure, liability, and product safety

Andrew F. Daughety and Jennifer F. Reinganum

1 INTRODUCTION

In this chapter we consider how models of imperfect competition, developed by scholars working in industrial organization (IO), provide insight into an important area of law: products liability (that is, liability for harms and losses associated with goods and services sold via markets). This importance derives from the fact that everyday life generally involves consumption activities wherein the risk of harm is present: we all consume manufactured goods and commercially harvested and/or prepared foods, translocate or telecommute between home and employment, and occupy space in buildings and homes that condition the air we breathe and the light we use (not to mention relying on the safety of those structures).

Remarkably, traditional law and economics (L&E) analyses of products liability generally find no role for the influence of market structure or strategic interaction on liability policy. Two results come from the traditional analysis. First, different liability regimes (to be detailed below) lead to the same private choices of safety, and this private choice is the socially optimal level of safety.¹ Second, alternative market structures (perfect competition, monopoly, oligopoly) have no effect on the level of safety chosen by firms. In what follows, after briefly summarizing the traditional analysis, we consider two simple (but plausible) model modifications that yield a substantial impact of market structure on the choice of safety and (potentially) on the choice of liability regime. Section 2 provides a summary of the traditional model used to consider unilateral precaution in the case of harms due to products; Section 3 then reconsiders the traditional model of harm while Section 4 reconsiders the traditional model of production cost. In both Sections 3 and 4 we first consider monopoly provision of the product and then extend the model to the case of oligopoly competition. Section 5 provides a brief review of additional contributions to this literature and summary comments.

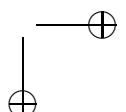
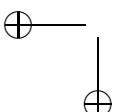
2 THE TRADITIONAL L&E MODEL OF PRODUCTS LIABILITY UNDER UNILATERAL PRECAUTION²

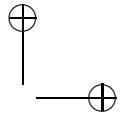
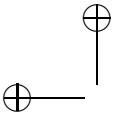
2.1 Preliminaries: Consumers

A representative consumer has quasilinear utility for two goods, the product of interest (consumed in an amount $q \geq 0$) and a “numeraire” good of amount $z \geq 0$; the consumer’s

¹ Formally, this statement assumes that precaution against an accident occurring is unilateral: only one agent’s choice of precaution (usually taken to be the manufacturer of the product) affects the expected harm. We focus on the unilateral precaution case in this chapter; for a discussion of the bilateral precaution case (i.e., wherein the consumer must also take care), see Shavell (1987) for the model in Section 2, and Daughety and Reinganum (2013b) for the model in Section 3.

² A comprehensive discussion of the traditional model can be found in Shavell (1987).





income is I and the numeraire's price will always be 1 while the price of the good of interest (abstracting from safety considerations) is $p \geq 0$.³ Let $U(q, z) = u(q) + z$ be the utility the consumer derives from consumption of (q, z) , again abstracting from any safety considerations associated with the good of interest. Let $h(x)$ be the expected harm (per unit of the good) that the consumer suffers when the firm supplying the product has taken care level x . Finally, depending upon the liability regime in place, consumers who are harmed may be compensated. Under *strict liability*, the consumer will be compensated by the firm for any harm due to use of the product (that is, independent of the level of precaution the firm took) while under no liability, the consumer is not compensated for harm by the firm.⁴ In what follows, let r denote the liability regime and let δ be an indicator variable wherein $\delta = 1$ if the liability regime is no liability ($r = NL$) while $\delta = 0$ if the liability regime is strict liability ($r = SL$). Thus, the consumer's total utility from consuming the bundle (q, z) is $U(q, z; x, \delta_r) = u(q) + z - \delta_r h(x)q$, for $r = SL, NL$.⁵

Assumption 1

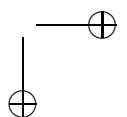
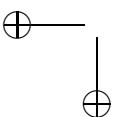
- (a) *The direct utility for the good of interest, $u(q)$, satisfies (for all $q \geq 0$):*
 - (i) $u(q) \geq 0$;
 - (ii) $u'(q) > 0$; and
 - (iii) $u''(q) < 0$.
- (b) *Per-unit expected harm, $h(x)$ satisfies (for all $x \geq 0$):*
 - (i) $h(0)$ is positive and finite, and $h(x) > 0$ for all $x > 0$;
 - (ii) $h'(x) < 0$; and
 - (iii) $h''(x) > 0$.
- (c) *The consumer's maximum willingness-to-pay for a unit of the good exceeds the worst per-unit expected harm: $u'(0) > h(0)$.*

Assumption 1(a) implies downward-sloping demand for the good of interest. Assumption 1(b) reflects the traditional assumption about investment in safety by the firm: increased investment reduces the expected harm, but there are diminishing returns to investment in safety with respect to reduction of the expected harm. Assumption 1(c) bounds the expected cost of harm; we assume that we are considering products that should not be removed from the market because they are inherently extremely dangerous or socially unacceptable. This is reasonable in a complete and perfect information model, since otherwise consumers could simply avoid the market.

³ In what follows, assume that income, I , is sufficiently large so that positive levels of both goods are consumed at optimality, and all functions are continuously differentiable as many times as needed.

⁴ A third alternative liability regime is negligence, wherein sufficient investment in safety meets a “due care standard,” relieving the firm of paying compensation, while insufficient investment implies “fault,” thereby requiring the firm to compensate the consumer. Thus, negligence is a hybrid of strict liability and no liability; we will address it later in this discussion. Under complete and perfect information, in the traditional model there would be no basis for a firm to take insufficient care, so it would never be found negligent (i.e., at fault). As will be discussed below, the modification of the model of harm to be discussed in Section 3 will actually predict that the firm may purposely violate the negligence standard, as it will prefer strict liability. For a recent discussion of these three liability regimes in the modeling context of products liability issues of manufacturing defects, design defects, and warning defects, see Daugherty and Reingenanum (2013a).

⁵ In what follows we focus on a complete and perfect information model, so as to direct attention to very basic changes in the analysis that arise from the assumptions made regarding harm and cost.



Notice that since strict liability ($\delta = 0$) is taken to imply perfect compensation of the consumer by the firm, then the consumer ignores the total expected harm, $h(x)q$, when buying q units of the good, whereas no liability means that the consumer will bear the entire cost of harm. To see this, solve the consumer's choice model, $\max_{q,z} U(q, z; x, \delta_r)$ subject to $pq + z \leq I$, which yields the (inverse) demand model for the good of interest:

$$p^r(q; x) = \max\{u'(q) - \delta_r h(x), 0\}, \text{ for } r = SL, NL. \quad (9.1)$$

Under assumption 1(c), this means that the first term in brackets above is always positive for some (q, x) combinations, in which case $p^r(q; x) = u'(q) - \delta_r h(x)$. That is, $p^{SL}(q; x) = u'(q) > 0$, and $p^{NL}(q; x) = u'(q) - h(x) > 0$. Given the earlier assumptions, this means that both demand functions are downward sloping, with $p^{NL}(q; x)$ downward shifted from $p^{SL}(q)$.

2.2 Preliminaries: Firms

Production reflects constant unit costs of production (this assumption will be modified in Section 4), though the unit cost of production is an increasing convex function of the level of safety.

Assumption 2 $C(x, q) = c(x)q$, with:

- (a) $c(0) = 0$ and $c(x) > 0$ for all $x > 0$;
- (b) $c'(x) > 0$ for all $x > 0$ and $c'(0) = 0$; and
- (c) $c''(x) > 0$ for all x .

In the rest of the discussion of the traditional model we assume that the firm is a monopolist; Shavell (1987, pp. 65–6) provides a discussion of the perfectly competitive case. Thus, the profit function for a firm operating under liability regime r can be written as:

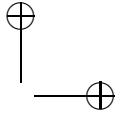
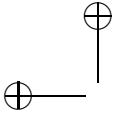
$$\Pi^r(q, x) = p^r(q, x)q - c(x)q - (1 - \delta_r)h(x)q. \quad (9.2)$$

That is, $\Pi^{SL}(q, x) = u'(q)q - c(x)q - h(x)q$, since under SL the firm fully compensates the consumer for any harm, while under NL , $\Pi^{NL}(q, x) = u'(q)q - h(x)q - c(x)q$, since the consumer discounts the value of the good by the expected cost of harm. Obviously, $\Pi^{SL}(q, x) = \Pi^{NL}(q, x)$, for all x and q ; that is, the consumer's willingness-to-pay fully reflects any anticipated uncompensated harm. Since second-order conditions are met, then the following first-order conditions characterize the firm's optimal choices in regime r , denoted as (q^r, x^r) :⁶

$$\Pi_x^r(q^r, x^r) = -(c'(x^r) + h'(x^r))q^r = 0; \quad (9.3)$$

$$\begin{aligned} \Pi_x^r(q^r, x^r) &= p^r(q^r, x^r) + p_q^r(q^r, x^r)q^r - c(x^r) - (1 - \delta_r)h(x^r) = u'(q^r) + u''(q^r)q^r \\ &\quad - (c(x^r) + h(x^r)) = 0. \end{aligned} \quad (9.4)$$

⁶ In the sequel, a subscript on a multivariate function indicates differentiation, e.g., $\Pi_x^r(x, q) \equiv \delta \Pi^r(x, q) / \delta x$. We use primes (' and '') for derivatives of single-variable functions.



The properties of $u(q)$, $c(x)$ and $h(x)$ imply that $q^r > 0$. Thus, from (9.3) one sees that x^r minimizes $c(x) + h(x)$ and is independent of both the equilibrium level of output and the liability regime: $x^{NL} = x^{SL}$. Note also that the simple (separable) structure of the first-order conditions (9.3) and (9.4) relies upon Assumption 2(a) (i.e., constant returns to scale in production of output) as well as the form of the expected harm function.

2.3 Efficiency and Primary Results for the Traditional Model

Welfare, denoted as $W(q, x)$, is given by:

$$W(q, x) = u(q) - c(x)q - h(x)q. \quad (9.5)$$

Assumptions 1 and 2 imply that $W(q, x)$ is strictly concave in (q, x) . Assuming a finite optimum, denoted as (q^W, x^W) , the first-order conditions require that (q^W, x^W) satisfy the following:

$$W_x(q^W, x^W) = -(c'(x^W) + h'(x^W))q^W = 0; \quad (9.6)$$

$$W_q(q^W, x^W) = u'(q^W) - (c(x^W) + h(x^W)) = 0. \quad (9.7)$$

Again, the properties in Assumptions 1 and 2 imply that q^W is positive and, through a comparison between equations (9.4) and (9.7), the usual simple monopoly result obtains: $q^r < q^W$; that is, the monopolist restricts output relative to the welfare-maximizing level.⁷ Notice also that equations (9.6) and (9.3) imply that $x^{NL} = x^{SL} = x^W$: the firm chooses the welfare-maximizing level of safety. Since $p^r(q; x) = u'(q) - \delta_r h(x)$, the separability in q and x means that $p_{qx}^r(q; x) = 0$. The foregoing welfare result is simply a reflection of the Spence (1975) condition concerning the under- or oversupply of quality; since the cross-derivative of the demand function is zero, the monopolist will choose to supply the efficient level of quality (here, safety). Again, from equation (9.6), this will be where the sum of the per-unit production cost and the per-unit expected loss from harm is minimized.

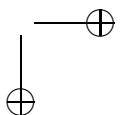
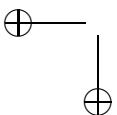
Proposition 1 summarizes the relevant results from the traditional model:

Proposition 1 *The traditional model (Assumptions 1 and 2) yields two results:*

- (a) *Firms (whether perfectly competitive or a monopoly) always choose to provide the efficient level of safety: $x^{NL} = x^{SL} = x^W$. Thus, the choice of liability regime does not matter.*
- (b) *Monopolists undersupply output, and this is unaffected by the liability regime chosen: $q^{NL} = q^{SL} < q^W$. Thus, market structure only affects the level of output provided.*

These results are readily extended to the traditional Cournot oligopoly version of the foregoing analysis. A qualifier for Proposition 1(a) is important to note. Since under no liability there will be no costs of formally adjudicating a “wrong,” while under strict liability some such

⁷ If the firms are price takers, then as Shavell shows (1987, pp. 65–6), both the output level and the safety level are efficiently provided by a competitive industry.



costs are bound to be incurred, some scholars have argued that NL is socially preferred to SL .⁸ We abstract from this argument about administrative costs in evaluating welfare.

A broad implication of the results from the traditional model is *policy separability*: antitrust and competition policies can be formulated and implemented without regard to the choice of tort regime in products liability: considerations of market performance are separable from considerations of product performance. Alternatively put, agencies or courts charged with formulating or implementing laws concerned with product market competition need not be particularly bothered with concerns about vertical product quality issues such as product safety, and the reverse as well: assessment of whether an agent is liable or culpable for harm does not require examination (by, say, a court) of the degree of competitiveness (or lack thereof) of the market wherein a good or service was acquired. This suggests an economy of governmental/judicial decision-making and it also suggests that the two areas of law can develop independently.

3 MODIFYING THE TRADITIONAL MODEL'S REPRESENTATION OF HARM: CUMULATIVE HARM

In the traditional model outlined in Section 2, total expected harm accrues in proportion to the quantity of the good consumed, so that marginal and average harm are the same. There is no obvious physical basis for this assumption; it is simply particularly useful in providing results as outlined in Section 2. In this section we modify this assumption and find that the resulting analysis provides significantly different results from those presented above.⁹ In particular, allowing for the model of total expected harm to be non-linear in the quantity of the good consumed means that the marginal and average harm are generally not the same; this implies that the choice of liability regime matters, and that market structure matters as well.

In Daughety and Reinganum (2014) (henceforth DR2014) we provide a number of motivating examples from food safety, pharmaceuticals, environmental safety, privacy (both on the Internet and in the physical world), and the operation of physical systems (i.e., wherein friction can have an effect on breakdown). In all of these examples, per-unit expected harm can be viewed as increasing in the quantity of the good consumed, so that total expected harm is actually convex in q . To capture this we modify Assumption 1 as follows:

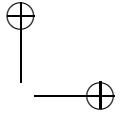
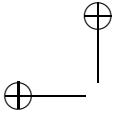
Assumption 1'

(a) The direct utility for the good of interest, $u(q)$, satisfies (for all $q \geq 0$):

- (i) $u(q) \geq 0$;
- (ii) $u'(q) > 0$; and
- (iii) $u''(q) < 0$.

⁸ For example, this is the underlying argument in Polinsky and Shavell (2010) for using NL in the case of mass-marketed products. Note that this argument relies strongly on the assumption that the choice of x by a firm (or firms) is perfectly observable to consumers, so that feedback via the market is effective. If x is unobservable, a firm under NL would pick a safety investment of zero and rational consumers should expect that $x^{NL} = 0$.

⁹ The results in this section draw heavily on Daughety and Reinganum (2014), hereafter denoted as DR2014, particularly Technical Appendix B. To our knowledge, the first paper to formally allow for and model non-linear (in q) expected harm is Marino (1988).



- (b) Total expected harm is modeled by the function $H(x, q)$, which is *thrice* continuously differentiable with the following properties:
- (i) $H(0, x) = 0$ for all $x \geq 0$; $H(q, x) > 0$ for all $q > 0$, and all $x \geq 0$;
 - (ii) $H_x(q, x) < 0$ and $H_{qq}(q, x) > 0$ for all $(q, x) > 0$; $H_x(0, x) = 0$ for all $x > 0$;
 - (iii) $H(q, x)$ is strictly convex in $(q, x) > 0$, with strictly negative off-diagonal terms; and
 - (iv) $H_{qqx}(q, x) < 0$ for all $(q, x) > 0$.
- (c) All optimization problems have a unique interior optimum, and all the respective payoff functions (W , Π^{SL} , and Π^{NL}) have second-order matrices that are negative definite in a sufficiently large neighborhood of these optima.

Assumption 1'(a) is the same as Assumption 1(a); the primary difference between Assumption 1' and Assumption 1 is in the total expected harm function, $H(q, x)$, which in Section 2 is $h(x)q$. Assumption 1'(b)(i) is parallel to Assumption 1(b). Assumptions 1'(b)(ii) and 1'(b)(iii) extend the convexity of $h(x)q$ used in Section 2 to now hold for both x and q in $H(q, x)$; $H_{xx} > 0$ parallels $h''(x) > 0$ from Section 2, but now we add $H_{qq} > 0$, which was absent from Section 2 (there $H_{qq} = 0$); intuitively, expected harm is decreasing in x but increasing in q at an exponential rate. Due to the multivariate nature of how harm is assumed to arise (i.e., both via the level of x and the level of q), we will need a further assumption about the interaction of q and x . There are diminishing returns to investment in safety ($H_{xx} > 0$), but increasing the investment in safety ameliorates the convexity of total expected harm in usage level, which requires that $\delta H_{qq}/\delta x < 0$.¹⁰ A handy example, consistent with all of the above properties of $H(q, x)$ is $h(x)q^2$, with $h(x)$ as in Section 2. This example was explored in the main text and in Technical Appendix A of DR2014; it was also exploited in the oligopoly discussion in Technical Appendix C of DR2014 and will be employed later in this section to review that oligopoly analysis.

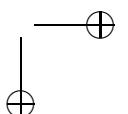
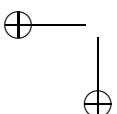
3.1 Monopoly Provision of Safety when Harm is Cumulative

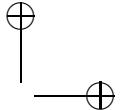
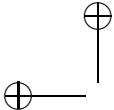
To provide the clearest insight, we again start with a monopolistic firm and derive an updated version of Proposition 1. In the subsequent subsection we proceed to consider the effect of allowing for oligopolistic competition (which can take one of two forms, depending upon the degree of interdependency of harm among the consumer's purchases). In a manner similar to that in Section 2, the (inverse) demand function for the good, conditional on the liability regime, is:

$$p^r(q; x) = u'(q) - \delta_r H_q(q, x), \quad \text{for } r = SL, NL, \quad (9.8)$$

which is positive for the relevant ranges of q and x .

¹⁰ As is shown in the Technical Appendix for DR2014, Assumption 1'(b)(iv) is equivalent to $H_{xq}(x, q) < H_x(x, q)/q$ for all $(x, q) > 0$. Marino (1988) instead considers the case wherein $H_{xq}(x, q) > H_x(x, q)/q$ for all $(x, q) > 0$; his focus is on "tolerance" wherein increased use of a good diminishes the rate of expected loss.





The convexity of H implies that the marginal expected harm exceeds the average expected harm:

$$H_q(q, x) > H(q, x)/q. \quad (9.9)$$

Equation (9.9) is key to understanding why a model with cumulative harm changes the results from those of the traditional model. To see this, we will specify the profit functions and first-order conditions for the *SL* firm (that is, the firm under a regime of strict liability), the *NL* firm (similarly, the firm under a regime of no liability), as well as the relevant welfare function and first-order conditions. Let a firm's profit under regime r again be denoted as $\Pi^r(q, x)$, where:

$$\Pi^r(q, x) = p^r(q, x)q - c(x)q - (1 - \delta_r)H(q, x). \quad (9.10)$$

Using equation (9.8), in the case of strict liability the *SL* firm's profit function ($\delta_{SL} = 0$) becomes:

$$\Pi^{SL}(q, x) = u'(q)q - c(x)q - H(q, x), \quad (9.11)$$

and the first-order conditions for the *SL* firm become:

$$\Pi_x^{SL}(q, x) = -c'(x)q - H_x(q, x) = 0; \quad (9.12)$$

$$\Pi_q^{SL}(q, x) = u'(q) + u''(q)q - c(x) - H_q(q, x) = 0. \quad (9.13)$$

Again, using equation (9.8), in the case of no liability, the *NL* firm's profit function ($\delta_{NL} = 1$) becomes:

$$\Pi^{NL}(q, x) = u'(q)q - c(x)q - H_q(q, x), \quad (9.14)$$

and the first-order-conditions for the *NL* firm become:

$$\Pi_x^{NL}(q, x) = -c'(x)q - H_{qx}(q, x)q = 0; \quad (9.15)$$

$$\Pi_q^{NL}(q, x) = u'(q) + u''(q)q - c(x) - H_{qq}(q, x)q - H_q(q, x) = 0. \quad (9.16)$$

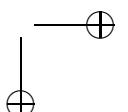
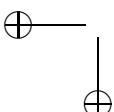
Finally, welfare is as follows:

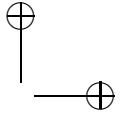
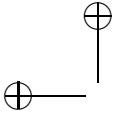
$$W(q, x) = u(q) - c(x)q - H(q, x), \quad (9.17)$$

and the first-order conditions for a maximum are the following:

$$W_x(q, x) = -c'(x)q - H_x(q, x) = 0; \quad (9.18)$$

$$W_q(q, x) = u'(q) - c(x) - H_q(q, x) = 0. \quad (9.19)$$





232 *Handbook of game theory and industrial organization: applications*

Some observations are now in order. First, as remarked upon earlier, the convexity of H in q implies that $H_q(q, x) > H(q, x)/q$ (the marginal expected harm due to increased consumption exceeds the average expected harm), so comparing equations (9.11) and (9.14) provides the immediate result that for any given relevant (q, x) :

$$\Pi^{NL}(q, x) < \Pi^{SL}(q, x).$$

That is, for given (q, x) , a firm will prefer a regime of strict liability to a regime of no liability; this reflects the fact that while in the traditional model no liability results in a downward shift of the demand curve from that under strict liability, under cumulative harm, no liability results in a rotation of the demand curve. For example, if $H(q, x) = h(x)q^2$, then demand under no liability becomes $u'(q) - 2h(x)q$ rather than what would obtain under strict liability: $u'(q)$.

Second, for any given q , equations (9.12) and (9.18) yield the same value of x ; that is, for a given output level, a firm under the *SL* regime provides the welfare-maximizing level of safety; a comparison of equations (9.15) and (9.18) show this does not hold for the *NL* regime. More precisely, let $x^{SL}(q)$ solve equation (9.12), let $x^{NL}(q)$ solve equation (9.15), and let $x^W(q)$ solve equation (9.18). Then evaluating equations (9.12) and (9.15) at $x^{SL}(q)$, and recalling the properties of H (see footnote 10) yields:

$$\Pi_x^{NL}(q, x^{SL}(q)) = H_x(q, x^{SL}(q)) - H_{qx}(q, x^{SL}(q))q > 0,$$

meaning that at the given level of output, q , the *NL* firm would choose a higher level of safety. That is:

$$x^{SL}(q) = x^W(q) < x^{NL}(q).$$

All three functions are upward sloping¹¹ and equal zero when the quantity level is zero. We can think of these as a type of (non-strategic) “best response” function, as they provide the optimal choice of the level of safety investment for any exogenously specified level of quantity.

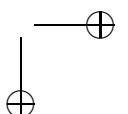
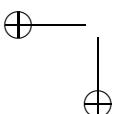
Let $q^{SL}(x)$ solve equation (9.13), let $q^{NL}(x)$ solve equation (9.16), and let $q^W(x)$ solve equation (9.19). These functions provide (respectively) the *SL*, *NL*, and *W* levels of optimal output for any arbitrary given level of safety investment, x . In a manner similar to that used above, note that:

$$\Pi_q^{NL}\left(q^{SL}(x), x\right) = -H_{qq}\left(q^{SL}(x), x\right)q^{SL}(x) < 0,$$

so that for any given x , $q^{NL}(x) < q^{SL}(x)$. Similarly,

$$\Pi_q^{SL}\left(q^W(x), x\right) = u''(q^W(x))q^W(x) < 0,$$

¹¹ For this and other details, see the Technical Appendix, Part B, for DR2014; in the sequel, we refer to this as TAB2014.



so the overall ordering for any given x is:

$$q^W(x) > q^{SL}(x) > q^{NL}(x).$$

Let $(\hat{x}^{SL}, \hat{q}^{SL})$ simultaneously solve equations (9.12) and (9.13); then $\hat{x}^{SL} = x^{SL}(\hat{q}^{SL})$ and $\hat{q}^{SL} = q^{SL}(\hat{x}^{SL})$. Similarly, let $(\hat{x}^{NL}, \hat{q}^{NL})$ simultaneously solve equations (9.15) and (9.16), so that $\hat{x}^{NL} = x^{NL}(\hat{q}^{NL})$ and $\hat{q}^{NL} = q^{NL}(\hat{x}^{NL})$, and finally let (\hat{x}^W, \hat{q}^W) simultaneously solve equations (9.18) and (9.19), so $\hat{x}^W = x^W(\hat{q}^W)$ and $\hat{q}^W = q^W(\hat{x}^W)$. In general $q^{SL}(x)$, $q^{NL}(x)$, and $q^W(x)$ are not monotonic, but one can show (see TAB2014) that these functions are increasing up to, and beyond, where the related “best response” safety functions (respectively $x^{SL}(q)$, $x^{NL}(q)$, and $x^W(q)$) cross the associated “best response” quantity functions, and that these crossings are from below. Figure 9.1 illustrates the foregoing for the special case wherein $H(q, x) = h(x)q^2$, where $h(x)$ is as in Section 2.

Two important lessons show up in Figure 9.1.¹² First, the policies *SL* and *NL* provide very different outcomes, in contrast with the traditional analysis from Section 2: market structure (as reflected by the level of output) really does matter now. The *NL* safety curve is to the right of the *SL* curve, since in order to reduce the effect on profits (which arises since the consumer discounts via the demand function) the *NL* firm increases x and simultaneously lowers q (which in turn means that the good’s price is higher in the market than would otherwise occur). Second, if the liability regime is *SL*, then increases in output brought about by, say, antitrust policy, results in $(x^{SL}(q), q)$ moving closer to (\hat{x}^W, \hat{q}^W) , while under an *NL* policy, $(x^{NL}(q), q)$ diverges from (\hat{x}^W, \hat{q}^W) .¹³ In other words, despite the interrelatedness of the safety and quantity decisions, the policy-separability property we discussed in Section 2 as a valuable result of the traditional model is available if agencies and courts involved in products liability tort actions employ strict liability: product performance (i.e., concerns about the safety of products) is still separable from market performance (i.e., concerns about the competitiveness of the marketplace), and decisions by agencies and courts, as well as the development of law regarding these two spheres, can proceed independently under *SL*.

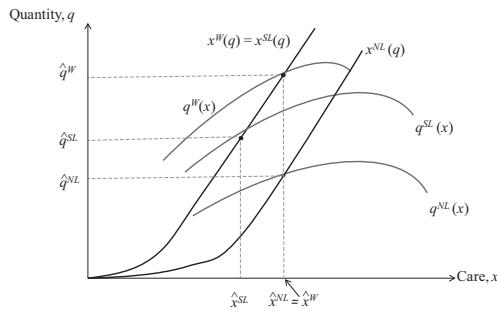
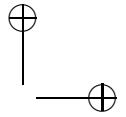
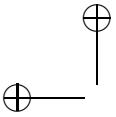


Figure 9.1 The cumulative harm results

¹² Figure 9.1 shows the equilibrium safety level under *NL* is the same as the welfare-optimal level; this is a result of the quadratic form employed, and is not a general property.

¹³ For the simple example wherein $H(q, x) = h(x)q^2$, welfare is always higher under *SL* when compared with *NL*; in the more general $H(q, x)$ case, this will be true in a small enough neighborhood of (\hat{x}^W, \hat{q}^W) .



One further point worth making is about how recognition of the non-proportionality of harm with respect to consumption affects policy. Earlier we purposely chose to ignore negligence. Negligence is a hybrid of *SL* and *NL*: a “due care” standard is set (perhaps at \hat{x}^W). A firm that chooses x below this standard is fully liable for all harm while a firm that chooses x to be at least this great faces no liability. Recall the result earlier that under cumulative harm, that $\Pi^{NL}(q, x) < \Pi^{SL}(q, x)$ for any given (q, x) . As we show in DR2014, this preference for *SL* by the firm means that it has an incentive to undermine a regime of negligence by purposely choosing x below the due care standard if consumers observe the firm’s choice of x prior to purchase (as has been assumed here). This would imply an *SL*-style market exchange, which involves stronger product demand than would occur under compliance with a negligence standard (resulting in *NL*).

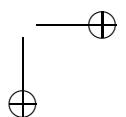
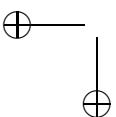
3.2 Oligopolistic Provision of Safety When Harm Is Cumulative

There are seemingly two means by which harm might accumulate via consumption of a collection of products. On the one hand, consuming products from n firms, each of which independently creates some harm, which is (say) convex in the amount of the product consumed, may be viewed as creating aggregate harm additively in the more-than-proportional consumption of each of the n products. In DR2014 we refer to this as the “Diner’s Dilemma” by imagining the risks at an open buffet of dinner items. One can get E. coli from a portion of beef, or salmonella from a portion of chicken, or shigella from a portion of vegetables. Upon becoming ill following dinner at the buffet, the particular source of harm (the product) can be identified. On the other hand, bioaccumulation of mercury can come about from consumption of a variety of mercury-containing fish.¹⁴ In this case the harm may be difficult to associate with any one product or source: harm is convex in the aggregate exposure to mercury that results from consumption of a variety of products.

3.2.1 Independent cumulative harm from n products

To see the influence of market considerations, first consider the case wherein the consumer views the n products as perfect substitutes in consumption (ignoring safety) and assume that product i ’s expected harm function is $h(x_i)(q_i)^2$, $i = 1, \dots, n$, wherein each h function has the same properties as the h function in Section 2 (i.e., decreasing but convex in x_i). Here we are assuming that harms are independent and the source of a harm is identifiable. To facilitate providing clear results, assume that the direct utility a consumer enjoys from consuming $\mathbf{q} = (q_1, q_2, \dots, q_n)$ of the n products is $u(\mathbf{q}) = \alpha \sum_i q_i - (\beta/2)(\sum_i q_i)^2$ where α and β are positive constants. Let $Q = \sum_i q_i$, so the (inverse) demand for product i is $\alpha - \beta Q$ when the liability regime is *SL* and $\alpha - \beta Q - 2h(x_i)q_i$ when the liability regime is *NL*. Therefore, under *SL*, firm i ’s profit function is $\Pi_i^{SL}(x_i, q_i; n) = (\alpha - \beta Q)q_i - h(x_i)(q_i)^2 - c(x_i)q_i$. We find the first-order conditions for x_i and q_i , respectively, and then invoke symmetry, yielding:

¹⁴ The US Food & Drug Administration provides a strong warning, especially for pregnant women, nursing women, and young children, concerning the consumption of the following types of fish: shark, swordfish, king mackerel, and tilefish, alone or in combination. See <https://www.fda.gov/Food/ResourcesForYou/Consumers/default.htm>, accessed September 14, 2017.



$$-h'(x)q^2 - c'(x)q = 0; \quad (9.20)$$

$$\alpha - (n+1)\beta q - 2h(x)q - c(x) = 0. \quad (9.21)$$

Simultaneous solution of equations (9.20) and (9.21) yields the symmetric Cournot-Nash equilibrium combination of care and output, denoted $\hat{x}^{SL}(n)$ and $\hat{q}^{SL}(n)$, respectively.

Similarly, an *NL* firm i 's profit function is $\Pi_i^{NL}(x_i, q_i; n) = (\alpha - \beta Q - 2h(x_i)q_i)q_i - c(x_i)q_i$. We find the first-order conditions for x_i and q_i , respectively, and then invoke symmetry, yielding:

$$-2h'(x)q^2 - c'(x)q = 0; \quad (9.22)$$

$$\alpha - (n+1)\beta q - 4h(x)q - c(x) = 0. \quad (9.23)$$

Simultaneous solution of equations (9.22) and (9.23) yields the symmetric Cournot-Nash equilibrium combination of care and output, denoted $\hat{x}^{NL}(n)$ and $\hat{q}^{NL}(n)$, respectively.

Finally, welfare when the consumption vector is \mathbf{q} and the safety vector is $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is:

$$W(\mathbf{q}, \mathbf{x}) = \alpha \sum_i q_i - (\beta/2)(\sum_i q_i)^2 - \sum_i h(x_i)(q_i)^2 - \sum_i c(x_i)q_i.$$

Differentiating to obtain the first-order conditions and then applying symmetry provide the following conditions for a welfare maximum:

$$W_x = -nh'(x)q^2 - nc'(x)q = 0; \quad (9.24)$$

$$W_q = n\alpha - \beta n^2 q - 2nh(x)q - nc(x) = 0. \quad (9.25)$$

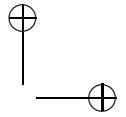
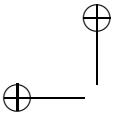
Let the solution to equation (9.20) be denoted as $x^{SL}(q; n)$, and let the solution to equation (9.24) be denoted as $x^W(q; n)$. Comparing equations (9.20) and (9.24) shows that:

$$x^{SL}(q; n) = x^W(q; n).$$

Moreover, modifying $u(q)$ and $H(q; x)$ to fit the current example, and comparing equation (9.12) with equation (9.20) and equation (9.24) with equation (9.18), shows that:

$$x^{SL}(q; n) = x^W(q; n) = x^{SL}(q) = x^W(q),$$

so that, for any given level of output q , the *SL* firm will choose the same level of safety for a given q independent of the number of firms, and this will always be the socially optimal safety choice for that level of output. Not surprisingly, $x^{NL}(q; n)$ is also independent of n and maintains the same relationship with $x^W(q)$ as it held in the monopoly analysis. However, the number of firms does affect the (symmetric) equilibrium quantity levels under *SL*, *NL*, or welfare maximization as specified in equations (9.21), (9.23), and (9.25), respectively; in DR2014 we show how the fact that firms undersupply output affects the equilibrium supply of safety (now no longer conditional on output level). In particular, we show that, as a function of the number of firms and computed at the equilibrium oligopoly output in each case, *SL*



firms undersupply safety while NL firms oversupply safety (i.e., $\hat{x}^{SL}(n) < \hat{x}^W(n) < \hat{x}^{NL}(n)$). Under SL , NL , and W , as the number of firms increases, each firm takes less care and produces a lower level of output, but total output increases and risks are increasingly diversified over a larger set of firms from which to purchase products.

3.2.2 Interdependent cumulative harm from n products

As pointed out earlier, the consumption of goods from different firms could lead to an overall accumulation of harm because the alternative goods each contribute to the extent of exposure to the same source of harm and it is the overall level of exposure that matters. A straightforward example is that provided earlier: consumption of different types of fish, all of which contain mercury. If the expected harm depends upon the aggregate level of exposure from all the goods, then the aggregate effect now is (potentially) cumulative. Thus, let $\mu(x_i)$ denote the amount of exposure per unit of good i consumed, where $\mu(x_i)$ has the same properties as $h(x_i)$ (that is, $\mu(0)$ is positive and finite, $\mu'(x) < 0$, and $\mu''(x) > 0$ for all x), so that $\mu(x_i)q_i$ represents the amount of exposure from consuming q_i units of product i . Thus, the aggregate amount of exposure is $\sum_i \mu(x_i)q_i$. We assume the expected harm is convex in the aggregate level of exposure; in DR2014 we consider the simple quadratic form of this:

$$(\sum_i \mu(x_i)q_i)^2. \quad (9.26)$$

Notice the contrast between this version of harm and that of the previous independent-harm model: $\sum_i h(x_i)(q_i)^2$.

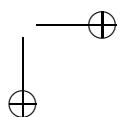
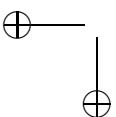
Space limitations prevent us from providing details of the analysis of this second model, but in DR2014 we show that, in contrast with the independent harms model, the total expected harm arising under an interdependent-harm process stays constant as n increases, so the risk of harm is not diversified as n increases. Thus, for example, if there are fixed costs of production, this means that society is better off restricting entry to one firm and regulating the product's price, or subsidizing production, to increase quantity. Second, unlike our results earlier, no standard liability regime provides the incentive to take the socially optimal level of care for a given level of output.

However, using our specific model of harm in equation (9.26), a simple multiplier in the firm's accounting for the expected harm, along with the use of exposure-based market-share liability (i.e., liability in proportion to $\mu(x_i)q_i / \sum_j \mu(x_j)q_j$), will make an SL firm choose the socially optimal level of safety for a given output level. The multiplier is $[2n/(n + 1)]$, so that firm i 's profit function under SL incorporates this multiplier times the firm's share of the exposure it contributes, times the total expected harm:

$$(\alpha - \beta Q)q_i - [(2n/(n + 1))\{\mu(x_i)q_i / \sum_j \mu(x_j)q_j\}]\{\sum_j \mu(x_j)q_j\}^2 - c(x_i)q_i. \quad (9.27)$$

Note that the firm bears additional liability (since $1 \leq 2n/(n + 1) < 2$); the extra liability payment should go to the state and not be part of compensation paid to harmed consumers. As shown in DR2014, this allocation of liability induces each SL firm's choice of safety to be the same as that which a central planner would choose (for a given output level).¹⁵

¹⁵ Market share allocations of harm have been used by courts to allocate liability across firms in an industry when the source of the harm (that is, which firm produced the product that caused a harm) was unclear; the most famous such case was *Sindell v. Abbott Laboratories*, 26 Cal. 3d 588 (1980).



4 MODIFYING THE TRADITIONAL MODEL'S PRODUCTION COST: SEQUENTIAL CHOICE OF SAFETY AND OUTPUT

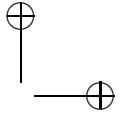
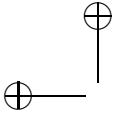
Another significant deviation from the basic products liability model involves a modification of how choices about investment in safety affect the cost function. In particular, many products with a safety attribute require upfront investment in R&D and product design. For example, investment in the development of airbags for automobiles must occur before any airbags are installed in automobiles and before those autos are subsequently sold. Investment in the development of pharmaceuticals must be done, including testing for safety and effectiveness, before any of the products reach the market. Thus, safety involves an endogenously determined fixed cost of development; it is also likely to impact the variable cost of production. Typically, we view such a fixed investment as a long-run choice whereas output can be varied in the short run. This suggests that firms' product design investments (which determine the safety of their products) should be viewed as being chosen in a prior period, with firms' output choices being made in subsequent periods. Since the safety levels are long-lived, it is also plausible to view them as being observed by all firms before output levels are chosen.

In Daughety and Reinganum (2006; hereafter DR2006), we provide a model wherein a number of firms compete non-cooperatively in a market for a product with safety attributes. In the first period, firms make investments in product design; a higher investment implies a safer product, but a safer product also has a higher marginal cost of production. Firms then observe the safety level of competitors' products (or, equivalently, they observe rival firms' fixed investment and can infer the implied safety level), and choose how much output to produce. Firms' products have some inherent level of horizontal differentiation, and they have at least the potential for vertical differentiation, since firms can choose any level of safety. Because the products are differentiated and consumers value variety, we model this as N identical consumers consuming some of each good. Consumption results in some accidents and harm to consumers. Although we will assume that firms are strictly liable for the harm their products cause, we also assume that the compensation process is imperfect so that consumers are left with some uncompensated harm.¹⁶ This could reflect tort reforms such as a cap on non-economic damages or the presence of litigation costs.¹⁷ As a consequence, a consumer's willingness-to-pay for a product will be sensitive to the product's safety level, as the consumer will anticipate having to bear some uncompensated loss should the product cause harm.

Some of the details of the DR2006 model will be simplified or suppressed here; see that paper for a full development and discussion of the model. We will use the following notation:

¹⁶ For simplicity, we will abstract from other issues that can introduce inefficiency. For example, we abstract from asymmetric information about the level of harm, which can result in negotiation failure and expenditure on litigation; we also abstract from externalities in the sense of harm caused to third parties. These issues are considered in DR2006, and the interested reader is referred to that paper for the details.

¹⁷ Even if a case settles in pretrial negotiation (so no actual litigation costs are spent), if the firm has substantial bargaining power then the consumer's settlement may not fully compensate her. For example, in the extreme case wherein the firm makes a take-it-or-leave-it offer, the settlement offer would equal the consumer's harm minus the costs she would incur by rejecting the offer and taking the case to trial.



238 *Handbook of game theory and industrial organization: applications*

- n the number of firms, each of which produces a variety of the good; the $n+1$ st good is a numeraire;
- N the number of identical consumers in the market;
- x_i firm i 's safety level;
- q_i firm i 's output level per consumer; thus, firm i 's total output is Nq_i ;
- t the unit cost of safety; thus, firm i 's total investment in safety is tx_i ;
- $m(x_i)$ the marginal cost of output for firm i ; thus, firm i 's variable cost of providing q_i units is $m(x_i)q_i$;
- $v^c(x_i)$ the portion of the expected harm borne by the consumer;
- $v^f(x_i)$ the portion of the expected harm borne by the firm.

We maintain the following assumptions about the functions $m(x_i)$ and the total expected harm $v^c(x_i) + v^f(x_i)$. The marginal cost of production, $m(x_i)$, is strictly positive, strictly increasing, and strictly convex in safety level. That is, marginal cost is constant with respect to output, but a safer product is more costly to produce. The total expected harm, $v^c(x_i) + v^f(x_i)$, encompasses multiple effects of safety. It reflects the probability of an accident; the extent of harm conditional on the occurrence of an accident; and any other costs such as litigation costs. Both the probability of an accident and the extent of harm conditional on the occurrence of an accident may be affected by the safety level. We assume that the total expected harm is strictly positive, strictly decreasing, and strictly convex. We revert here to the assumption that expected harm is proportional to the level of consumption (as in the base model in Section 2, and in contrast to the model with cumulative harm in Section 3), but we assume that safer products generate lower expected harm.

We assume that each consumer has a quasilinear utility function, $U(\mathbf{q}, z)$, where \mathbf{q} is again the n -vector of quantities consumed and z is the numeraire good, so that:

$$U(\mathbf{q}, z) = \sum_i \alpha q_i - (1/2)[\sum_i \beta(q_i)^2 + \sum_i \sum_{j \neq i} \gamma q_i q_j] + z. \quad (9.28)$$

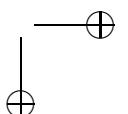
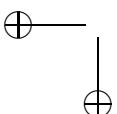
The parameters α , β , and γ are positive, with $\gamma \leq \beta$. That is, products 1 through n are imperfect substitutes. The parameter γ represents the extent of horizontal product differentiation; as γ converges to β , the products become perfect substitutes (in terms of horizontal differentiation).

There are two ways to incorporate the consumer's expected uncompensated harm, $v^c(x_i)$. Given that it is assumed to reflect a constant risk per unit of the good consumed, one way is to view it as a reduction in utility. In this interpretation, we would substitute the expression $\alpha - v^c(x_i)$ for the parameter α in the utility function. Alternatively, we could view this expected loss in dollar terms and incorporate it into the consumer's budget constraint. Letting I denote the consumer's income, budget balance requires that $\sum_i [(p_i + v^c(x_i))q_i] + z = I$. Either of these interpretations results in the same formal optimization problem for the consumer.

$$\max_{\mathbf{q}} \sum_i \alpha q_i - (1/2)[\sum_i \beta(q_i)^2 + \sum_i \sum_{j \neq i} \gamma q_i q_j] + I - \sum_i [(p_i + v^c(x_i))q_i].$$

The resulting inverse demand function for product i is given by:

$$p_i(\mathbf{q}; x_i) = \alpha - v^c(x_i) - \beta q_i - \sum_{j \neq i} \gamma q_j. \quad (9.29)$$



Firm i 's profit function is then given by:

$$\pi_i(\mathbf{q}; x_i) = Nq_i[p_i(\mathbf{q}; x_i) - m(x_i) - v^f(x_i)] - tx_i.$$

That is, firm i sells Nq_i units overall, and makes a profit of $p_i(\mathbf{q}; x_i) - m(x_i) - v^f(x_i)$ on each unit. The safety investment, tx_i , increases with the safety level x_i , but it does not vary with output. Upon substituting for the function $p_i(\mathbf{q}; x_i)$ and collecting terms, we can rewrite firm i 's profit as:

$$\pi_i(\mathbf{q}; x_i) = Nq_i[\alpha - \beta q_i - \sum_{j \neq i} \gamma q_j - FMC(x_i)] - tx_i, \quad (9.30)$$

where "full marginal cost" $FMC(x_i) = m(x_i) + v^c(x_i) + v^f(x_i)$. That is, although the expected losses from harm are nominally shared by the consumer and the firm in the amounts of $v^c(x_i)$ and $v^f(x_i)$, respectively, the consumer simply reduces her willingness-to-pay for product i by the amount of her expected uncompensated harm as shown in equation (9.29). Thus, the firm faces the full marginal cost (that is, the marginal production cost and the entire expected loss from accidents per unit of output).

In addition to the previous assumptions we made regarding the individual functions $m(x_i)$, $v^c(x_i)$, and $v^f(x_i)$, we further assume that $FMC(x_i)$ is convex and "U-shaped" so there is a safety level that minimizes full marginal cost. Denote this safety level by \bar{x} ; this is the level of safety that would be chosen in the traditional law and economics model of product safety wherein there is no endogenous fixed cost reflecting, say, safety design (i.e., wherein $t = 0$). It will become clear in what follows that a firm's safety level will always be less than \bar{x} . Finally, we assume that $FMC(0) < \alpha$ (so the product can always be produced profitably).

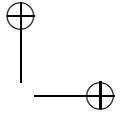
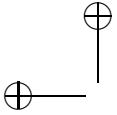
We will use first-order conditions to characterize the firms' equilibrium output and safety levels. Conditional on the vector of safety levels, the firms' profit functions are strictly concave in their respective output levels and the subgame perfect equilibrium in output levels can be computed directly. Stepping back to the simultaneous choice of safety levels, this concavity is no longer guaranteed. Nevertheless, we will assume that the reduced-form profit functions (as a function of the vector of safety levels, taking into account how these affect the subgame perfect output levels) are strictly concave in the firms' respective safety levels and that the overall (symmetric) equilibrium is characterized by the first-order conditions. We will use superscripts to indicate the number of firms; for instance, x^1 and q^1 will denote a monopolist's profit-maximizing choice of safety and output.

For the case of monopoly, $\pi(q; x) = Nq[\alpha - \beta q - FMC(x)] - tx$. The profit-maximizing output level, conditional on the safety level x , is $q^1(x) = (\alpha - FMC(x))/2\beta$. Substituting this into the profit function yields the reduced-form profit as a function of safety level: $\Pi(x) = N(\alpha - FMC(x))^2/4\beta - tx$. Thus, the monopolist's profit-maximizing choice of safety level is given by:

$$(N/2\beta)[\alpha - FMC(x^1)][-FMC'(x^1)] - t = 0,$$

or, equivalently:

$$-Nq^1(x^1)FMC'(x^1) - t = 0. \quad (9.31)$$



It is clear that $x^1 < \bar{x}$; that is, because the unit cost of safety t is positive, the firm will choose a safety level less than that which minimizes its full marginal cost (and therefore $FMC'(x^1) < 0$). It is straightforward to demonstrate the following comparative statics results. The monopolist's safety level x^1 (and output level q^1) increases with an increase in α or N , both of which generate an increase in the size of the market for the product (a higher α reflects an increase in an individual consumer's willingness-to-pay for the product and a higher N reflects a larger number of consumers). The monopolist's safety level x^1 (and output level q^1) decreases with an increase in t (the unit cost of safety) or β (the rate at which a consumer's willingness-to-pay for the product declines with the quantity consumed).

Now we consider the oligopoly version of the model wherein n firms first non-cooperatively choose safety levels and subsequently, having observed rival safety levels, the n firms non-cooperatively choose output. The profit function has been provided above in equation (9.30). The first-order condition for firm i 's choice of output is:

$$\alpha - 2\beta q_i - \sum_{j \neq i} \gamma q_j - FMC(x_i) = 0.$$

Solving the collection of first-order conditions for the subgame perfect vector of output level choices (conditional on the vector of safety level choices, denoted \mathbf{x}) yields:

$$q_i^n(\mathbf{x}) = [(2\beta - \gamma)\alpha - (2\beta + (n-2)\gamma)FMC(x_i) + \gamma \sum_{j \neq i} FMC(x_j)] / [(2\beta - \gamma)(2\beta + (n-1)\gamma)]. \quad (9.32)$$

It is clear that each firm's output level is increasing in its own safety level and decreasing in its rivals' safety levels.

Now consider firm i 's choice of safety level, anticipating how it will affect all firms' subgame perfect choices of output. The first-order condition for x_i is given by:

$$\{-Nq_i^n(\mathbf{x})FMC'(x_i) - t\} + Nq_i^n(\mathbf{x})\gamma \sum_{j \neq i} [-\delta q_j^n(\mathbf{x})/\delta x_i] = 0. \quad (9.33)$$

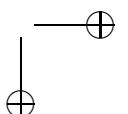
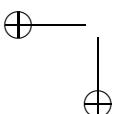
The term in brackets is similar to that in equation (9.31) and represents the simple trade-off between a higher fixed cost of safety and a lower full marginal cost of output. The second term is positive since $\delta q_j^n(\mathbf{x})/\delta x_i = \gamma FMC'(x_i)] / [(2\beta - \gamma)(2\beta + (n-1)\gamma)] < 0$. An increase in firm i 's safety level reduces its rivals' subgame perfect output levels, which translates into a higher inverse demand curve for firm i . This "business-stealing" effect, which is absent from the monopoly model, provides an additional marginal benefit from raising one's safety level.

In what follows, we denote the symmetric equilibrium safety level for an n -firm oligopoly by x^n , and we denote the symmetric equilibrium output level for an n -firm oligopoly by q^n . Under symmetry, equations (9.32) and (9.33) become:

$$q^n(x^n) = [\alpha - FMC(x^n)] / [2\beta + (n-1)\gamma]; \quad (9.34)$$

$$-Nq^n(x^n)FMC'(x^n)\{1 + \gamma^2(n-1)/[(2\beta - \gamma)(2\beta + (n-1)\gamma)]\} - t = 0. \quad (9.35)$$

The comparative statics results that were derived earlier in the monopoly case extend to the case of oligopoly: the safety level x^n (and output level q^n) increases with an increase in α or N , and decreases with an increase in t or β . However, now two additional parameters



appear in the model. Both n and γ are measures of market competitiveness; clearly a higher number of competitors (n) makes competition more intense, but so does a greater degree of substitution (γ).

An increase in the number of firms results in a lower equilibrium safety level x^n and output level q^n . The intuition for this is that, all else equal, more firms in the market will lead to a higher overall output but a lower output per firm. A firm that anticipates a lower output will find it optimal to choose a lower safety investment; although this investment lowers full marginal cost, this cost reduction applies to fewer units of output. Thus the return to investment in safety is lower when the number of firms is higher.

An increase in the degree of substitution has more complex effects. Recall that $\gamma \in [0, \beta]$. When $\gamma = 0$ each firm is a monopolist, whereas when $\gamma = \beta$ the products are perfect substitutes. In general, the degree of horizontal differentiation has a direct effect on the consumer's marginal willingness-to-pay for each good (a higher value of γ means a lower marginal willingness-to-pay for good i) and it has an indirect effect on how responsive firm j 's output level is to the safety level of good i (a higher value of γ makes business stealing more effective). When γ is relatively small the business-stealing effect of an increase in firm i 's safety level on firm j 's output is also relatively small as the products are poor substitutes. Thus the overall effect is that x^n decreases as γ increases, when γ is small. But when γ is sufficiently large, then the indirect (business-stealing) effect dominates the direct effect (that is, the reduced marginal willingness-to-pay for any one variety of the product) and x^n increases as γ increases. More formally, there is a threshold value $\gamma^{\min}(\beta, n) \in [0, \beta]$ such that x^n decreases as γ increases for $\gamma < \gamma^{\min}(\beta, n)$ and x^n increases as γ increases for $\gamma > \gamma^{\min}(\beta, n)$. Moreover, it can be shown that $\gamma^{\min}(\beta, n)$ is increasing in n and that $\lim_{n \rightarrow \infty} \gamma^{\min}(\beta, n) = \beta$. The impact of increasing γ on q^n is also complex; it can be shown that q^n decreases as γ increases for $\gamma \leq \gamma^{\min}(\beta, n)$, but the effect of further increases in γ is ambiguous.

The dashed curves in Figure 9.2 illustrate the effects of increasing market competitiveness through either increasing n or increasing γ . The extent of horizontal product differentiation (γ) is measured along the horizontal axis. The equilibrium safety level (x^n) is measured along the vertical axis. For $n = 1$ (or for any n , if $\gamma = 0$), there is a single value of the safety level, denoted x^1 ; this is less than \bar{x} , which is the safety level at which FMC is minimized. For any particular $n > 1$, the equilibrium safety level first falls as γ increases, but then rises once γ exceeds $\gamma^{\min}(\beta, n)$. As n increases, each curve depicting x^n is everywhere below those for smaller values of n (except that they start at the same point, x^1).

We now consider how a social planner that is interested in promoting socially optimal safety might behave. However, we will not consider an all-powerful social planner (that could, in principle, choose the number of firms, the safety levels, and the output levels). Rather, our social planner will take the number of firms and their non-cooperative behavior with respect to output choices as given. Under these circumstances, what safety level would the social planner choose? If this is different from what the firms themselves would choose in equilibrium, can the liability system be modified or augmented with other policies to achieve the desired safety level?

Given a common safety level, denoted X (which will be chosen by the planner), the market operates as before. Consumers determine their willingness-to-pay for the products and the firms non-cooperatively choose their output levels. This generates utility for consumers, which our planner takes into account, and production and expected liability costs for the firms,

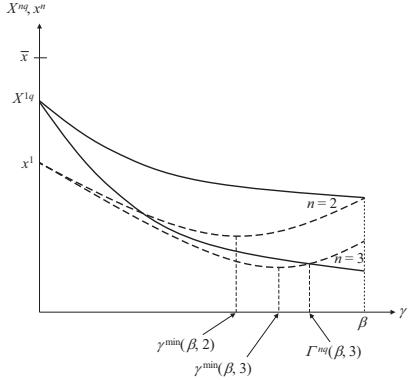


Figure 9.2 Market vs socially efficient choice of care with product differentiation

which our planner also takes into account. Since our planner takes n as given and firms non-cooperatively choose their output levels, subgame perfect equilibrium output will be given by the same function as before. That is,

$$q^n(X) = [\alpha - FMC(X)]/[2\beta + (n-1)\gamma]. \quad (9.36)$$

The planner's problem is then to choose X to maximize:

$$NU(q_1 = \dots = q_n = q^n(X)) - n[FMC(X)Nq^n(X) + tX]. \quad (9.37)$$

Let X^{nq} denote the planner's optimal choice of safety level (we use the superscript “ nq ” to indicate that the number of firms and the non-cooperatively chosen output levels are taken as given by the planner). The resulting first-order condition is:

$$-Nq^n(X^{nq})FMC'(X^{nq})\{(3\beta + (n-1)\gamma)/(2\beta + (n-1)\gamma)\} - t = 0. \quad (9.38)$$

Since the firms choose output non-cooperatively in both scenarios, these are only different to the extent that the social planner's safety level (X^{nq}) is different from the non-cooperative safety level (x^n). The comparative statics effects of X^{nq} with respect to increases in n , N , α , and β are the same as those of x^n . However, whereas x^n was first decreasing and then increasing with an increase in γ , the social planner's safety level X^{nq} always decreases with an increase in γ . This is because there is no social return to business stealing. Indeed, it can be shown that there exists a threshold value $\Gamma^{nq}(\beta, n) \in [0, \beta]$ such that non-cooperative firms would choose a lower safety level than the planner for $\gamma < \Gamma^{nq}(\beta, n)$, and a higher safety level than the planner for $\gamma > \Gamma^{nq}(\beta, n)$. For $n = 2$, it turns out that $\Gamma^{nq}(\beta, n) = \beta$; however, since $\Gamma^{nq}(\beta, n)$ decreases as n increases, it follows that $\Gamma^{nq}(\beta, n) < \beta$ for $n > 2$.

The solid curves in Figure 9.2 illustrate the effects on X^{nq} of increasing market competitiveness through either increasing n or increasing γ . Increasing either measure of competitiveness (holding the other one fixed) results in a lower value of X^{nq} . Furthermore, the figure illustrates the result that when competition is sufficiently intense due to a substantial number of firms

and/or a high degree of substitution, the non-cooperative firms' equilibrium safety level can exceed what a social planner would choose (given non-cooperative output choice).

One reason that we have considered a social planner that can only choose safety (and not output, or the number of firms) is that this social planner might be analogous to a court whose concern is whether or not a firm has provided appropriately safe products. In individual (or class action) products liability suits, the court does not assess all aspects of how the market functions (e.g., does each firm produce the socially optimal amount of output and is the number of firms socially optimal?), but rather it focuses on whether the particular firm before the court provided an appropriately safe product. In what follows, we will ask how the liability system can be modified to induce firms to choose the socially optimal safety level, taking as given the number of firms and how the firms will choose to supply output (i.e., as Cournot-Nash players).

Since the firms produce according to the usual oligopoly formula in both scenarios, the easiest way to induce non-cooperative firms to choose the (constrained) socially optimal level of care, X^{nq} , is to impose a penalty on the non-cooperative firms for deviating from X^{nq} . This could be a proportional penalty, such as $\lambda(X^{nq} - x_i)$, where λ is a constant, or simply a large fixed penalty that is imposed whenever firm i chooses x_i not equal to X^{nq} . This penalty would be paid only once by each firm, and it would be paid to the state. It could be interpreted as a penalty for negligence or an assessment of punitive damages; the firms would remain strictly liable to consumers. Notice that we are speaking as if inefficient safety decisions would always be revealed. But since each firm produces Nq units, the likelihood that at least one unit causes harm (which triggers the litigation that verifies the firm's true safety choice, enabling the one-time imposition of punitive damages) is very close to one, at least for mass-marketed products.

The specific value of λ that is required to induce non-cooperative firms to choose $x^n = X^{nq}$ can be found easily. Let $\theta^n = \{1 + \gamma^2(n - 1)/[(2\beta - \gamma)(2\beta + (n - 1)\gamma)]\}$ and let $\Theta^{nq} = \{(3\beta + (n - 1)\gamma)/(2\beta + (n - 1)\gamma)\}$; these are, respectively, the terms in braces in equations (9.35) and (9.38). Then the non-cooperative firms' equilibrium safety level occurs where $-Nq^n(x^n)FMC'(x^n) = (t - \lambda)/\theta^n$, whereas the socially optimal safety level occurs where $-Nq^n(X^{nq})FMC'(X^{nq}) = t/\Theta^{nq}$. To induce these safety levels to coincide, we simply need: $(t - \lambda)/\theta^n = t/\Theta^{nq}$, or $\lambda = t(1 - \theta^n/\Theta^{nq})$.

It is straightforward to show that $x^n < X^{nq}$ when $\theta^n < \Theta^{nq}$; that is, the penalty rate λ is positive when non-cooperative firms would undersupply safety absent the penalty. On the other hand, $x^n > X^{nq}$ when $\theta^n > \Theta^{nq}$; if the non-cooperative firms would oversupply safety absent the penalty, then the penalty rate λ must be negative (so that the overall penalty, $\lambda(X^{nq} - x^n)$, is positive). It is somewhat incongruous to think of courts imposing punitive damages for products that are "too safe," but in this situation business-stealing incentives have led the firms to compete too aggressively in terms of safety (driving up the endogenously determined fixed cost of safety).

5 OTHER MODELS OF IMPERFECT COMPETITION AND PRODUCT-GENERATED HARMS

In this section we describe a selection of particularly relevant other papers that involve models of imperfect competition in which firms choose safety and output (or price), and in which

consumers may be harmed by the product. There are many models involving competitive firms, and many models that do not involve markets at all (e.g., models of driving accidents), wherein liability is included as part of the model. Due to space constraints, we are not able to include a discussion of all of these related models.

Spence (1977) and Polinsky and Rogerson (1983) examine the impact of consumer misperceptions of safety on firms' choices of safety and output. To the extent that products liability leaves consumers undercompensated, they will deduct their (possibly misperceived) expected losses from their willingness to pay for the product, whereas the firm deducts its actual expected losses. Spence (1977) examines a competitive model with homogeneous goods wherein consumers underestimate the expected loss. He shows that when consumers are risk-neutral, the first-best outcome can be achieved (in terms of safety and output) by employing strict liability with full compensation.¹⁸ Polinsky and Rogerson (1983) examine an oligopoly model with homogeneous goods wherein consumers always underestimate the expected loss. They examine strict liability (with full consumer compensation), negligence, and no liability. Under strict liability, firms face the full expected liability costs; given that this is a proportional harm model, non-cooperative firms (that choose safety and output at the same time) always choose the socially optimal safety level but provide too little output. Under negligence, firms meet the negligence standard (which is set at the socially optimal level of safety) but all expected losses are borne by consumers. Since consumers underestimate the expected loss, they do not reduce their willingness to pay by the true expected loss. Thus, equilibrium output is higher under negligence than under strict liability, which is a welfare improvement. Finally, a regime of no liability has the same impact on consumer willingness to pay as negligence, but it does not support the socially optimal level of safety. Rather, firms now provide too little safety, which also lowers their full marginal costs, and causes even higher equilibrium output than under negligence. Polinsky and Rogerson argue that it may be optimal to take advantage of consumer misperceptions (of this specific type) by employing a liability regime that shifts the expected losses to consumers (either negligence or even possibly no liability, depending on the welfare tradeoff between lower safety and higher output).

Baniak, Grazl, and Guse (2014) describe a different sort of consumer misperception in an oligopolistic market; they assume that consumers cannot identify the safety levels of individual firms, but they are able to assess the average safety of products in the market (i.e., the firms share a collective reputation). The model is otherwise similar to that of Polinsky and Rogerson (1983), except that they allow the damages award to be arbitrarily different from the harm; that is, consumers could anticipate under- or overcompensation in the case of an accident. They compare the levels of safety provided in the regimes of strict liability and no liability. No liability results in too little safety, as improving own safety is a public good that is enjoyed by other firms through the collective reputation. Moreover, strict liability always results in higher equilibrium safety than does no liability. However, the authors show that when the damages exceed the harm, then strict liability can result in excessive safety. In this case, neither regime is obviously better: no liability provides too little safety whereas strict

¹⁸ Spence goes on to consider risk-averse consumers; he shows that the first-best can be achieved using "two-part liability," wherein the firm makes a payment to both the consumer and to the state in the event of an accident (the payment to the state compensates for consumers' underestimation of expected losses). He also considers whether voluntary liability can serve as a commitment to choosing higher safety.

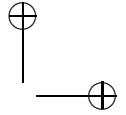
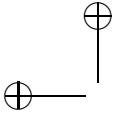
liability provides too much safety. Collective reputation also results in an interdependency between safety and output; the typical ranking is that output is lower under no liability than under strict liability, which is itself lower than the socially optimal output level. However, in the case of damages that are sufficiently in excess of the consumer's loss, strict liability can increase safety (and thus firm costs) greatly, which also greatly suppresses output; it is possible that no liability results in greater output than strict liability.

Daughety and Reinganum (1995) provide a monopoly model wherein the firm first makes an investment in safety (in particular, it engages in sequential search for its product design), and then sells the product to consumers; technically, the investment determines the type space from which the safety level will be drawn. Thus, the safety level is an exogenous attribute at the point of sale of the product; moreover, the safety level is the firm's private information. Although the safety level is unobservable to consumers prior to purchase, a firm's price can signal its product's safety. A safer product is assumed to have higher marginal production costs but lower marginal expected liability costs. We find that when firm liability is low (resp., high), the firm searches for a version with lower (resp., higher) production costs and higher (resp., lower) expected liability costs. Moreover, when firm liability is low the firm signals a higher safety level with a higher (than full-information) price that increases as safety increases, whereas when firm liability is high the firm signals a higher safety level with a lower (than full-information) price that decreases as safety increases. Since a monopoly provides too little output, these results suggest that the firm should bear a significant share of liability in order to induce lower price, greater output, and more search for safer designs.

Baumann and Friehe (2010) provide a monopoly model wherein a firm may have a high or low cost of providing safety (this is the firm's type). The firm produces over two periods and consumers cannot observe the firm's choice of safety in the first period prior to purchase. However, they do observe the firm's first-period safety level after first-period consumption and prior to second-period purchase. Thus, a firm can use first-period safety choice to signal something about its type (which will govern its choice of second-period safety as well). In particular, a low-cost type is willing to invest more than it would under observable safety, in order to deter mimicry by the high-cost type. The authors observe that this signaling motive results in higher consumer welfare than would occur if safety were observable prior to purchase. The authors do not, however, consider the potential for price to signal safety. Although the firm is a monopolist, it is not viewed as quoting a price for the good; rather, the price is simply taken as the consumer's maximum willingness-to-pay based on her conjectures about what safety level the firm has chosen.

Baumann and Friehe (2015) consider an oligopoly model wherein safety is an investment. However, they assume that this choice is not observable by other firms or by consumers. Rather, consumers have rational expectations about product safety. As a consequence, consumer willingness-to-pay is not sensitive to actual firm safety choices (it depends on consumers' conjectures about firm safety choices); moreover, the authors do not consider any inferences the consumer might draw from observing a firm's output (or price) choice. In this sense, the consumer may not be fully rational.¹⁹ A share of the expected harm is

¹⁹ In a standard signaling model, the firm's private information (its "type") is exogenous, and the firm's choices can reveal its type. In this alternative conjectures-based model, the firm's private information is its endogenous choice of safety level; nevertheless, there are models and solution concepts (e.g., forward induction) wherein the firm's public action can reveal its private action (see, for example, Dana, 2001). Thus, in order to incorporate full consumer rationality, it would be more appropriate for consumer conjectures about safety to be contingent on the firm's other



246 *Handbook of game theory and industrial organization: applications*

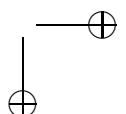
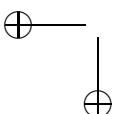
allocated to the firm, with the residual being borne by the consumer. However, they allow the firm's "share" to exceed 1, meaning that the firm pays the consumer more than her actual harm in the event of an accident. They then show that a social planner whose only instrument is a multiplier on the firm's liability to the consumer will choose a multiplier equal to $1 + 1/(n + 1)$, where n is the number of firms. Firms underinvest in safety in this model because consumers and rival firms cannot observe safety (thus removing the business-stealing motive). Unobservable safety is also why "scaling up" damages can work; consumer willingness-to-pay is affected by the scale factor but not by the firm's actual safety choice, whereas the firm itself recognizes that the scale factor generates a greater marginal impact of its safety choice on its expected liability payments to consumers. If consumers observed the firm's safety choice, then scaling damages up or down would have no effect on firm profits; any undercompensation (resp., overcompensation) would simply be deducted from (resp., added to) the consumer's willingness-to-pay for the product.

Finally, Ganuza, Gomez, and Robles (2016) consider a monopoly model wherein consumers are unable to observe a firm's choice of safety directly, but can retaliate in future periods following an incident of harm. In particular, in every period (of an infinite horizon), the firm chooses either a high or a low safety level. A safer product is more expensive, but less likely to cause harm, than a less-safe product. If the consumer could observe the safety level prior to purchase, she would buy the product only if safety was high. The authors characterize the following type of equilibrium (first taking the price as given, and subsequently with an endogenous price): the consumer starts out buying the product but, following an accident, she does not buy again for a number of periods. This number of periods (the "reputational penalty") is just sufficient to deter the firm from choosing low safety. The authors then incorporate liability in a flexible form that encompasses both strict liability (the firm is liable whenever an accident occurs) and negligence (the firm is liable when an accident occurs only if it chose low safety). They find that, in both cases, liability reduces the reputational penalty required to sustain high safety, but negligence requires an even lower reputational penalty than strict liability because it generates a finding regarding the firm's actual safety choice, rather than relying only on the occurrence of an accident (which is informative about the safety choice, but not perfectly so).

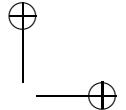
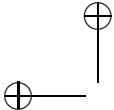
REFERENCES

- Baniak, A., P. Grajzl, and A.J. Guse (2014), "Producer liability and competition policy when firms are bound by a common industry reputation," *B.E. Journal of Economic Analysis and Policy*, **14**, 1645–76.
- Baumann, F. and T. Friehe (2010), "Product liability and the virtues of asymmetric information," *Journal of Economics*, **100**, 19–32.
- Baumann, F. and T. Friehe (2015), "Optimal damages multipliers in oligopolistic markets," *Journal of Institutional and Theoretical Economics*, **171**, 622–40.
- Dana, J.D. (2001), "Competition in price and availability when availability is unobservable," *RAND Journal of Economics*, **32**, 497–513.
- Daughety, A.F. and J.F. Reinganum (1995), "Product safety: Liability, R&D and signaling," *American Economic Review*, **85**, 1187–206.
- Daughety, A.F. and J.F. Reinganum (2006), "Markets, torts, and social inefficiency," *RAND Journal of Economics*, **37**, 300–323.

observable choice (output, or price), and then to explore whether this would allow consumers to draw inferences about the firm's unobservable choice (safety).



- Daughety, A.F. and J.F. Reinganum (2013a), "Economic analysis of products liability: Theory," in J. Arlen (ed.), *Research Handbook on the Economics of Torts*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 69–96.
- Daughety, A.F., and J.F. Reinganum (2013b), "Cumulative harm, products liability, and bilateral care," *American Law and Economics Review*, **15**, 409–42.
- Daughety, A.F. and J.F. Reinganum (2014), "Cumulative harm and resilient liability rules for product markets," *The Journal of Law, Economics, & Organization*, **30**, 371–400.
- Ganuza, J.-J., F. Gomez, and M. Robles (2016), "Product liability versus reputation," *Journal of Law, Economics, and Organizations*, **32**, 213–41.
- Marino, A.M. (1988), "Monopoly, liability, and regulation," *Southern Economic Journal*, **54**, 913–27.
- Polinsky, A.M. and W.P. Rogerson (1983), "Products liability, consumer misperceptions, and market power," *Bell Journal of Economics*, **14**, 582–9.
- Polinsky, A.M. and S. Shavell (2010), "The uneasy case for product liability," *Harvard Law Review*, **123**, 1437–93.
- Shavell, S. (1987), *Economic Analysis of Accident Law*, Cambridge, MA: Harvard University Press.
- Spence, A.M. (1975), "Monopoly, quality, and regulation," *Bell Journal of Economics*, **6**, 417–29.
- Spence, A.M. (1977), "Consumer misperceptions, product failure and producer liability," *Review of Economic Studies*, **44**, 561–2.



10. Strategic delegation in oligopoly*

Michael Kopel and Mario Pezzino

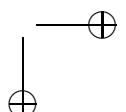
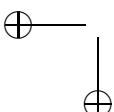
1 INTRODUCTION

In his book *Strategies of Commitment*, Nobel Prize winner Thomas Schelling discusses several examples of the functioning of commitment strategies and their importance (Schelling, 2006). The idea is that in an environment of strategic interdependence where each player's payoff depends on the decisions of all players, a player's credible commitment to a strategy is important as it shapes rivals' expectations and thus influences outcomes. For example, when Airbus publicly announced its plans to build the super jumbo jet A380, competitor Boeing dropped the plans to pursue a higher-capacity version of the successful 747 (called the 747X). Arguably, the visible commitment of Airbus has preempted Boeing to further compete in the market for super jumbo jets, a market that is believed to be too small to be profitable for two large-airframe manufacturers.

Strategic delegation is an example of *strategic commitment*. Delegation of decision making is common. Shareholders delegate managerial decisions to experienced chief executive officers (CEOs). Headquarters delegate investment and price or quantity decisions to division managers with local market know-how. Manufacturers delegate marketing decisions to their retailers who are well informed about customer preferences. The main message of the strategic delegation literature is that delegation of decisions to agents with differing objectives can provide benefits to the owners by influencing the expectations and actions of rival firms. This is particularly important in oligopolistic markets where only a limited number of firms compete and the performance of each firm in the market depends on the choices of all firms. In oligopolistic markets, observable transfer of decision rights to agents who are making the choices on behalf of the firms' principals can yield strategic benefits. For example, delegation allows the firm's owners to contractually induce the manager to compete more aggressively by selling a higher quantity in the product market than would be sold without delegation. To exploit the benefits from delegation, the principal can strategically influence the delegate's choices by properly designing compensation contracts, by limiting decision rights, or by adapting the organizational structure. The strategic delegation approach complements the theory of incentives in principal–agent relationships, where the design of contracts or organizational structures serves to minimize the costs of asymmetric information and conflicting interests of the parties.

One objective of this chapter is to provide the reader with a clear and intuitive, but yet rigorous, description of the topic of strategic managerial incentives under oligopolistic competition. A review of the closely related issue of vertical separation where a manufacturer delegates decisions to a retailer and an agent appointment game where a principal delegates decisions to a certain type of agent is also provided. We start the discussion

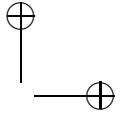
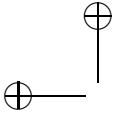
* This chapter has benefited from the insightful comments and suggestions of Harald Hinterecker, Matthias Kräkel, Clemens Löffler, Anna Ressi, and Christian Riegler.



of each of these themes by reflecting on seminal papers that have first introduced the corresponding topic. We present and critically discuss the key assumptions behind each of the basic models and point out important applications along with some empirical and experimental evidence. We also discuss contributions that have provided important extensions to the basic frameworks. Our chapter can be seen as complementary to the extensive literature on agency and control issues. Agency theory focuses, e.g., on the optimal design of the agent's compensation contract under asymmetric information. While it addresses the costs that arise because of firm-internal frictions due to asymmetric information and opportunistic behavior, it commonly does not take into account strategic interactions in the product market. In contrast, our chapter focuses on strategic delegation, i.e., contractual or organizational solutions as a device for influencing competition at the product or input market. In this literature, the frictions between the principal and agent might actually have strategic benefits. We hope that this chapter is helpful for readers who possess only a rudimentary understanding of game theory and are interested in acquiring a solid and intuitive understanding of the contributions of the literature of strategic delegation. We also hope that more experienced readers find some inspiration in our discussions of the more subtle points of the theory of strategic delegation and its extensions of the standard models. It is worth mentioning that we do not intend to provide an exhaustive review of the vast literature on strategic delegation. There are some (but few) reviews available. Gal-Or (1997) summarizes the early literature on the topic and discusses its limitations and extensions. Irmens (1998) concentrates on the literature that studies strategic aspects of vertical separation and the design of distribution channels. Sengul, Gimeno, and Dial (2012) provide a more comprehensive review of the strategic delegation literature from a management research perspective.

The chapter is structured as follows. Section 2 focuses on the strategic delegation framework, its assumptions, and its limitations. Section 2.1 introduces strategic incentives in a game with two owner-manager pairs and describes the key mechanism behind the basic model of strategic delegation. Specifically, we introduce a simple Cournot duopoly model where firm owners can delegate decision making (i.e., quantity choices) to managers. Each owner offers a publicly observable take-it-or-leave-it contract to its manager where compensation is based on profit and sales revenue. Section 2.2 critically discusses the assumptions of the basic model and mentions contributions that challenge these assumptions. We discuss the impact of changing the mode of competition from quantity to price competition (Section 2.2.1), relate the concept of strategic delegation to the broader idea of material and behavior payoffs (Section 2.2.2), and discuss the importance of asymmetric information between the owner and the manager (Section 2.2.3). We further take a closer look at alternative managerial contracts and performance measures such as market shares and relative performance contracts (Section 2.2.4), the importance of the assumption that managerial contracts are publicly observable (Section 2.2.5), and the possibility of owner-manager bargaining (Section 2.2.6). We end this section by providing empirical and experimental literature that supports the predictions of the strategic delegation framework (Section 2.2.7).

Strategic delegation is not limited to the relationship between owners of the firm and its managers. Vertical separation, i.e., the use of independent retailers and distributors, is another form of strategic delegation. Section 3 describes in detail a price-setting oligopoly game of competing vertical hierarchies when manufacturers have the choice of delegating the price decision to retailers that are, in turn, affected by the manufacturers' choices of the wholesale



prices. In particular, in Section 3.1 we derive closed-form solutions when the market is described by a Hotelling line and firms' locations are fixed. In Section 3.2 we critically discuss the key assumptions of the model. Specifically, we focus on linear wholesale contracts and alternative forms of vertical restraints (Section 3.2.1), on the assumption of observability of contracts and alternatives to the assumption of passive beliefs (Section 3.2.2), and the strategic design of distribution channels (Section 3.2.3). We further reflect on using transfer prices as a collusion device (Section 3.2.4) and briefly discuss bargaining between manufacturer and retailer over aspects of the wholesale contract (Section 3.2.5).

The objective of Section 4 is two-fold. First, we consider an agent appointment game and argue that the firm can effectively commit to a certain product market strategy by hiring a particular type of manager. In Section 4.1 we discuss literature that addresses the benefits and costs of such a strategy. Second, we abstract from downstream competition and argue that delegation (to a particular manager type) might have important effects on the upstream channel by influencing supplier behavior. In Section 4.2 we find that a (multi-product) monopolist can benefit from hiring a socially concerned manager who donates a certain amount of the sale revenues to charity as this commitment softens supplier pricing. Section 4.3 emphasizes the interplay between organizational design and the firm's supply side.

Finally, in Section 5 we present applications of the strategic delegation model to various important topics like R&D, horizontal mergers, and the endogenous emergence of firm heterogeneity. We end the chapter with a brief summary of the key points and some thoughts on future research opportunities.

2 DELEGATION AND INCENTIVE CONTRACTING

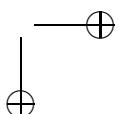
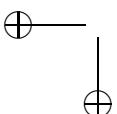
We begin our study of the effects of strategic delegation by considering a simple quantity-setting duopoly. To start with, we present the main ingredients of the model and some derivations, but postpone a more thorough discussion of the assumptions and extensions to later in this section.

2.1 Strategic Delegation and Stackelberg Quantity Leadership

Assume that market demand is characterized by the linear function $p(Q) = a - bQ$, where $Q = q_1 + q_2$ is the industry output, which is the sum of the quantities selected by the two firms. The choke price a and the slope b are both non-negative. Firm i 's cost of producing q_i units is given by $C_i(q_i) = c_i q_i$, where $a > c_i$ for $i = 1, 2$. We assume that marginal costs c_i are such that quantities in equilibrium are non-negative. Firm i 's profit function and revenue function are then given by, respectively,

$$\begin{aligned}\pi_i(q_1, q_2) &= (a - b(q_1 + q_2) - c_i)q_i, \\ R_i(q_1, q_2) &= (a - b(q_1 + q_2))q_i.\end{aligned}$$

The owners of the two firms are free to hire managers and delegate the output decision to them (see Basu, 1995). Both parties, owners and managers, are assumed to be risk-neutral in order to have a clear focus on possible strategic effects of delegating decisions. If a manager



is hired, the owner provides (strategic) incentives. That is, the owner of firm i offers manager i a (publicly observable) compensation contract based on profit and sales revenue,

$$w_i(q_1, q_2) = A_i + B_i [\alpha_i \pi_i + (1 - \alpha_i) R_i].$$

Here, $A_i \in \mathbb{R}$ denotes a fixed payment (or salary) and $B_i > 0$ denotes a bonus rate. The expression in the brackets, a weighted average of profit and sales revenue, is the performance measure used in manager i 's contract. The parameter $\alpha_i \geq 0$ represents the incentive rate set by the owner in manager i 's contract. Note that $\alpha_i = 0$ implies that the manager's performance is evaluated only in terms of revenues. Instead $\alpha_i = 1$ describes the case in which the manager's compensation is based only on profits. Manager i 's reservation income is denoted by $U_i \geq 0$.

The timing of our multi-stage game is as follows. First, owners either hire a manager ($m_i = 1$) or not ($m_i = 0$). If firm i does not hire a manager, the quantity is chosen by the owner of firm i to maximize profits π_i . If firm i hires a manager, then, in stage two, the owner of firm i determines the contract parameters A_i, B_i , and α_i such that the net profit, $\pi_i - w_i$, is maximized subject to the participation constraint $w_i \geq U_i$. Obviously, A_i and B_i will be adjusted so that in equilibrium the participation constraint binds and the net profit becomes $\pi_i - U_i$. The contract parameter α_i is used to provide (strategic) incentives to manager i . Finally, the quantity q_i is chosen by the manager such that compensation w_i is maximized.

The four subgames for all possible choices in period 1 can be easily solved by backward induction. To determine the overall outcome of the game, the normal-form game depicted in Table 10.1 can be used. The entries in the cells of the payoff matrix correspond to the owners' payoffs in each of the four subgames. We now elaborate on these subgames and the resulting payoffs.

Both owners do not delegate If both owners do not delegate ($m_1 = m_2 = 0$), then the game corresponds to a standard quantity-setting ("Cournot") duopoly. We use the superscript N to express (Cournot-Nash) equilibrium values when owners decide not to hire a manager. The owners' best reply (or reaction) functions are given by

$$q_i(q_j) = \max \left[0, \frac{a - c_i}{2b} - \frac{1}{2} q_j \right]$$

and are represented graphically by the continuous lines in Figure 10.1. The resulting quantities are $q_i^N = \frac{a - 2c_i + c_j}{3b}$ (indicated by point E_1 in Figure 10.1) and the profits are given by $\pi_i^N = bq_i^2 = \frac{(a - 2c_i + c_j)^2}{9b}, i, j = 1, 2 (i \neq j)$.

Table 10.1 Normal-form game to determine the overall equilibrium of the game

		Owner 2	
		$m_2 = 0$	$m_2 = 1$
Owner 1	$m_1 = 0$	π_1^N, π_2^N	$\pi_1^F, \pi_2^L - U_2$
	$m_1 = 1$	$\pi_1^L - U_1, \pi_2^F$	$\pi_1^D - U_1, \pi_2^D - U_2$

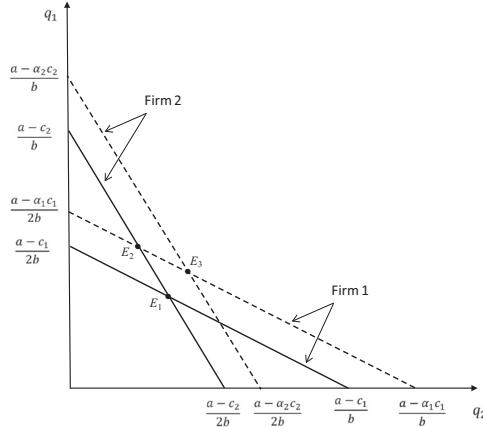


Figure 10.1 Best reply functions under Cournot duopoly with delegation (dashed lines) and without delegation (continuous lines)

Only one owner delegates Assume without loss of generality that only the owner of firm 1 hires a manager, i.e., $m_1 = 1, m_2 = 0$. The other asymmetric case is obtained by swapping firm indices. In the final stage of this game, manager 1 chooses the production quantity q_1 such that this manager's compensation w_1 is maximized. Note that manager 1's compensation can be rewritten as $w_1 = A_1 + B_1[R_1 - \alpha_1 c_1 q_1]$. Hence, in contrast to the owner, manager 1 acts as if the marginal costs of production are not c_1 but $\alpha_1 c_1$. By choosing the incentive rate α_1 , e.g., smaller than 1, the manager's perceived marginal production costs are lower than the actual marginal costs. Consequently, the owner can effectively turn the manager into a more aggressive player in the quantity selection stage of the game. For firm 2 the owner is choosing the quantity q_2 so that profit π_2 is maximized. Formally, from the first-order conditions manager 1's best reply

$$q_1(q_2) = \max \left[0, \frac{a - \alpha_1 c_1}{2b} - \frac{1}{2} q_2 \right]$$

and the rival firm 2's best reply

$$q_2(q_1) = \max \left[0, \frac{a - c_2}{2b} - \frac{1}{2} q_1 \right]$$

can be derived. Solving yields similar expressions for the quantities as before, with marginal costs of firm 1 replaced by $\alpha_1 c_1$,

$$q_1(\alpha_1) = \frac{a - 2\alpha_1 c_1 + c_2}{3b}$$

$$q_2(\alpha_1) = \frac{a - 2c_2 + \alpha_1 c_1}{3b}.$$

Note that both quantities depend on manager 1's incentive rate α_1 .

The reduced-form profits can be written as $\pi_i(q_1(\alpha_1), q_2(\alpha_1))$. The owner of firm 1 now determines the incentive rate α_1 to maximize profits. One might be tempted to conclude that the “right” incentives are provided by $\alpha_1 = 1$, since in this case the owner of firm 1 offers the manager a contract that provides incentives to maximize profits, replicating the owner-managed outcome from above. However, it turns out that the owner can do better than that. Note that by selecting $\alpha_1 < 1$, manager 1’s best reply is shifted outwards (see the dashed line of firm 1 in Figure 10.1). The manager then would select a higher quantity and the owner of the rival firm a smaller quantity (point E_2 in Figure 10.1). In our case, quantities are strategic substitutes (Bulow, Geanakoplos, and Klemperer, 1985). The owner of firm 1 can strategically provide incentives to its manager to become more aggressive and steal market share from its rival. As a consequence, this increases firm 1’s profit. This can be easily seen by differentiating the reduced-form profit of firm 1 and using the envelope theorem,

$$\underbrace{\frac{\partial \pi_1}{\partial q_1} \frac{\partial q_1}{\partial \alpha_1}}_{=0} + \underbrace{\frac{\partial \pi_1}{\partial q_2} \frac{\partial q_2}{\partial \alpha_1}}_{<0 >0},$$

which shows that decreasing the incentive rate α_1 increases firm 1’s profit. Delegation is an example of a so-called “top dog” strategy (see Fudenberg and Tirole, 1984). Likewise, we have

$$\underbrace{\frac{\partial \pi_2}{\partial q_1} \frac{\partial q_1}{\partial \alpha_1}}_{<0 <0} + \underbrace{\frac{\partial \pi_2}{\partial q_2} \frac{\partial q_2}{\partial \alpha_1}}_{=0},$$

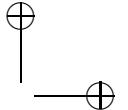
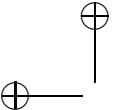
which demonstrates that firm 2’s profit shrinks with decreasing α_1 . What is the optimal choice of the incentive parameter α_1 and the induced quantity q_1 ? To understand this note that firm 2’s best reply $q_2(q_1)$ does not depend on α_1 , while the best reply $q_1(q_2)$ of firm 1’s manager shifts outwards for decreasing α_1 . In other words, by selecting α_1 , firm 1’s owner can (indirectly) select an intersection point along firm 2’s best reply by manipulating its manager’s best reply. This is akin to the problem of a Stackelberg leader firm that (directly) selects its quantity given the Stackelberg follower’s best reply. Consequently, firm 1’s owner will choose the incentive rate α_1 , which will induce the manager to select the Stackelberg leader’s quantity (denoted by superscript L). Solving the first-order condition $d\pi_1(\alpha_1)/d\alpha_1 = 0$ yields

$$\alpha_1^L = 1 - \frac{a - 2c_1 + c_2}{4c_1}.$$

As anticipated, the associated quantities and profits coincide with the Stackelberg leader’s and follower’s quantities (superscript F) and profits,

$$q_1^L = \frac{a - 2c_1 + c_2}{2b}, \quad q_2^F = \frac{a - 3c_2 + 2c_1}{4b}, \\ \pi_1^L = \frac{(a - 2c_1 + c_2)^2}{8b}, \quad \pi_2^F = \frac{(a - 3c_2 + 2c_1)^2}{16b}.$$

Note that $\alpha_1 < 1$ if and only if $q_1^L > 0$. This is in line with the intuition provided above.



Both owners delegate The (sub)game where both owners hire managers ($m_1 = m_2 = 1$) has been studied in detail by Fershtman and Judd (1987) and Sklivas (1987). We briefly summarize the most important aspects of their derivation. For the final stage of this game where manager i chooses the production quantity q_i we obtain best replies (described graphically by the two dashed lines in Figure 10.1)

$$q_i(q_j) = \max \left[0, \frac{a - \alpha_i c_i}{2b} - \frac{1}{2} q_j \right].$$

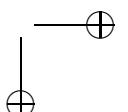
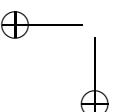
Consequently, given the contract for the manager of firm j , the owner of firm i is selecting the incentive parameter α_i to induce the manager to select the Stackelberg quantity. Since both owners give their managers incentives to increase their quantities (i.e., both best replies are shifted outwards), the industry output increases and the market price decreases. This situation resembles the case of a Stackelberg warfare and, as a consequence, both firms achieve smaller profits than in the owner-managed case. The subgame-perfect outcome of the game (point E_3 in Figure 10.1) where both owners delegate (superscript D) is

$$\begin{aligned}\alpha_i^D &= 1 - \frac{a - 3c_i + 2c_j}{5c_i} < 1, \\ q_i^D &= \frac{2(a - 3c_i + 2c_j)}{5b}, \\ \pi_i^D &= \frac{2(a - 3c_i + 2c_j)^2}{25b}.\end{aligned}$$

Having derived the entries of the payoff matrix given in Table 10.1, we are now able to determine the conditions for strategic delegation to be beneficial for a firm. For simplicity, we restrict our discussion to the case $U_1 = U_2 = U$. First note that for $U < \min[\pi_1^D - \pi_1^F, \pi_2^D - \pi_2^F] = \min[\frac{7(a-3c_1+2c_2)^2}{400b}, \frac{7(a-3c_2+2c_1)^2}{400b}]$ hiring a manager is an equilibrium. If, in addition, $U < \min[\pi_1^L - \pi_1^N, \pi_2^L - \pi_2^N] = \min[\frac{(a-2c_1+c_2)^2}{72b}, \frac{(a-2c_2+c_1)^2}{72b}]$ holds, then hiring is a dominant strategy for both owners. If, on the other hand, $\frac{7(a-3c_2+2c_1)^2}{400b} < U < \frac{(a-2c_1+c_2)^2}{72b}$, then firm 1 hires a manager while firm 2's owner does not. Under these conditions, strategic delegation of the quantity choice to a manager who is compensated by an appropriately chosen incentive contract endogenously gives firm 1 the Stackelberg leadership position. Note that regularly the manager's reservation utility is set to zero ($U = 0$), so that this asymmetric case does not occur and both firms delegate. In fact, the two firms are in a prisoner's dilemma and delegation is a dominant strategy. Due to the fact that both firms induce their managers to choose quantities that are higher than under pure profit maximization, both firms end up with lower profits.

2.2 Assumptions of the Model and Extensions

The model presented in the previous subsection has a number of restrictions and limitations that we are going to discuss at length in this subsection. We further discuss extensions of the original model and the robustness of the results if some assumptions are relaxed.



2.2.1 Oligopolistic markets and mode of competition

Fershtman and Judd (1987) consider the quantity-setting oligopoly case with $n > 2$ firms and demonstrate that for $n \rightarrow \infty$, firms would give incentives for maximizing profits. They conclude that the relation between managerial incentives and the number of firms is non-monotonic. They also consider a differentiated-products duopoly where managers choose prices. They demonstrate that incentives given to managers are very different than under quantity competition. In fact, managers are punished for sales. They are given (strategic) incentives to act less aggressively, i.e., to keep prices high. Since prices are strategic complements, delegation in this case works as a commitment mechanism to price less aggressively, which in turn leads to higher profits than in the case of owner management for both firms. Similarly, Sklivas (1987) studies strategic delegation based on profit and sales revenue under quantity choice and price choice with identical conclusions.

To provide some useful intuition at the outset, let us consider in some detail a simple Bertrand duopoly where goods are imperfect substitutes. In Section 3 we will study in more detail a model of price competition under vertical separation where a franchisor delegates the price choice to a franchisee. Consider a market served by two firms, say firms 1 and 2, providing goods that are imperfect substitutes. Firms face the following (inverse) demands, $p_i = 1 - q_i - \gamma q_j, i = 1, 2, j \neq i$. Parameter γ represents the degree of substitutability between goods produced by the two firms. If $\gamma = 0$ then the two firms effectively are two monopolists operating in different markets. If $\gamma \rightarrow 1$, then goods would be perfect substitutes. In what follows we assume that $\gamma = \frac{1}{2}$ to increase the tractability of the equilibrium expressions. Direct demands faced by the two firms are $q_i = \frac{2}{3}(1 + p_j - 2p_i)$. Firm i 's cost of producing q_i units is given by $C_i(q_i) = cq_i$, where $0 \leq c < 1$.

Firm i 's profit function and revenue function are then given by, respectively,

$$\begin{aligned}\pi_i(p_1, p_2) &= (p_i - c) \left(\frac{2}{3}(1 + p_j - 2p_i) \right) \\ R_i(p_1, p_2) &= p_i \left(\frac{2}{3}(1 + p_j - 2p_i) \right).\end{aligned}$$

Similar to the Cournot duopoly studied above, we assume that the (risk-neutral) owners of the two firms are free to hire (risk-neutral) managers and delegate the price decision to them. If a manager is hired, the owner provides (strategic) incentives. That is, the owner of firm i offers manager i a (publicly observable) compensation contract based on profit and sales revenue,

$$w_i(p_1, p_2) = A_i + B_i[\alpha_i \pi_i + (1 - \alpha_i)R_i].$$

The timing of our multi-stage game is the same as in the Cournot example. To keep the analysis to a minimum, we will discuss only the two symmetric scenarios in which either both firms do not delegate or they both hire a manager. The asymmetric case will only be briefly discussed.

Both owners do not delegate If both owners do not delegate, then the game corresponds to a standard price-setting (“Bertrand”) duopoly with imperfect substitutes. The best reply functions of the two firms are

$$p_i(p_j) = \frac{(1 + 2c + p_j)}{4},$$

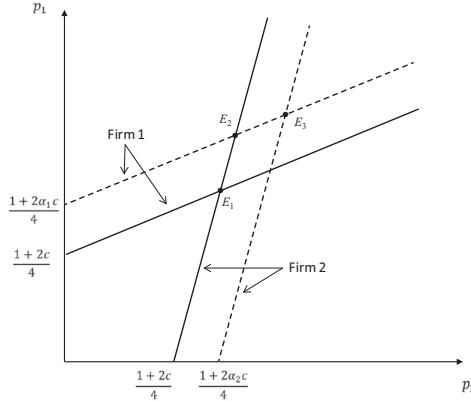


Figure 10.2 Best reply functions under Bertrand duopoly with delegation (dashed lines) and without delegation (continuous lines)

and described by the two continuous lines in Figure 10.2. Note that – in contrast to the best reply functions from Cournot competition – the best reply functions in prices are upward sloping, i.e., prices are strategic complements. The equilibrium prices (point E_1 in Figure 10.2) are $p_i^N = \frac{1+2c}{3}$ and the profits are given by $\pi_i^N = \frac{4(1-c)^2}{27}, i, j = 1, 2 (i \neq j)$.

Both owners delegate The analysis of the case in which only firm 1 delegates is similar to our analysis of the Cournot duopoly. We leave it to the reader to check that the price best reply of firm 1 shifts upwards (see the resulting dashed line of firm 1 and the equilibrium prices as given by point E_2 in Figure 10.2). In this case, the owner of firm 1 would choose the manager's incentive rate α_1 such that its firm becomes the Stackelberg price leader. Let us now consider the case in which both owners hire managers. Price best replies are given by

$$p_i(p_j) = \frac{1 + p_j + 2\alpha_i c}{4},$$

and are depicted graphically by the two dashed lines in Figure 10.2. The subgame-perfect outcome of this game is

$$\begin{aligned} \alpha_i^D &= \frac{21 + \frac{1}{c}}{22} > 1, \\ p_i^D &= \frac{4 + 7c}{11}, \\ \pi_i^D &= \frac{56(1-c)^2}{363}. \end{aligned}$$

For simplicity, we restrict our discussion to the case $U_1 = U_2 = U = 0$. In this case, delegation is a dominant strategy. Note that the equilibrium incentive rate is larger than 1. This means that the manager is punished for sales. Consequently, prices are set less aggressively (i.e., higher) than under owner management. The managers' best replies are shifted outwards

and the market prices increase (point E_3 in Figure 10.2). As a consequence, both firms achieve higher profits than in the owner-managed (no delegation) case, $\pi_i^D > \pi_i^N$.

2.2.2 Material versus behavior payoffs

The key to strategic delegation is the principal's ability to credibly commit to transfer important decision rights to a delegate and control the delegate's incentive by a publicly observable contract. As a consequence, the strategic interdependence with other players might allow the principal to earn higher payoffs. If the principal were able to credibly commit to such an induced strategy without delegation of decision making, then the effects would nonetheless be comparable to those reported by the strategic delegation literature. Conceptually, strategic delegation of decision rights to a manager can be related to a broader principle whereby one needs to distinguish between material payoffs (the true utility function that expresses the payoff that an economic agent enjoys) and behavior payoffs (the utility used by the economic agent when determining strategic actions). The “inner” self of an individual may be selfish and targets the maximization of material payoffs, but the agent can delegate decision making to an “outer” self that may target instead the maximization of a behavioral payoff. Rotemberg (1994) argues that altruism can be used as a strategic tool that may allow economic agents to achieve higher payoffs. In other words, choosing to be altruistic and therefore considering the other's utility when choosing actions can be in the individual's self-interest. Of course, each agent's degree of altruism needs to be credibly demonstrated in order to have an effect on the behavior of others. Specifically, Rotemberg (1994) gives a number of examples where altruism increases material payoffs, for example in workplace scenarios in which altruism toward other employees could favor strategic complementarity and can be beneficial for an individual. He further discusses the endogenous choice of the optimal degree of altruism and reflects on long-run evolutionary outcomes if altruism can be beneficial. Casadesus-Masanell (2004) provides an application to a principal–agent setting under norms and ethical standards. Other papers further study the relation between behavioral and material payoffs. Koçkesen, Ok, and Sethi (2000) consider supermodular and submodular games that are symmetric with respect to material payoffs. Under certain conditions, players with interdependent preferences, i.e., players care about their payoffs relative to the payoff of other players, earn strictly higher material payoffs than selfish players. As an application they discuss strategic delegation and show that the results obtained for the linear model qualitatively carry over to a setting of non-linear demand and cost functions. Furthermore, they argue that under payoff-monotone selection dynamics (where better-performing strategies have a higher rate of replication in the evolutionary process) in the long run players with interdependent preferences will survive. Like an individual who could strategically decide to behave altruistically, firms can strategically commit to be socially responsible. Essentially, a firm could strategically decide to target the maximization of a combination of own profits and, for example, consumer surplus. If commitment to social responsibility is credible, then caring for customers could be a profitable strategy (see Königstein and Müller, 2001, Kopel, Lamantia, and Szidarovszky, 2014 and Kopel and Lamantia, 2016). Finally, Heifetz, Shannon, and Spiegel (2007) show that in almost every game and for almost every family of distortions of a player's material payoffs, some degree of this distortion leads to a higher material payoff. This behavior will survive under any evolutionary process based on payoff-monotone selection dynamics. They also provide an example where evolutionary dynamics does not lead to payoff-maximizing

behavior even if preferences are only imperfectly observed. As a consequence, the result that firms in oligopolistic markets can benefit from delegating decision rights to managers whose preferences differ from the owner's preferences is more generic than the simple linear model suggests.

2.2.3 Asymmetric information between the owner and the manager

An important issue often overlooked in contributions that employ the strategic delegation framework is information asymmetry between the owner and the manager. Fershtman and Judd (1987) are very clear about this (p. 930):

Uncertainty is also crucial to our focus on equilibria in which incentives are distorted away from profit maximization. We will argue that if we had no uncertainty about the ex post state of the market, then our analysis would be unconvincing since there would be no justification for ignoring quantity- or price-indexed contracts that would force the usual Cournot and Bertrand outcomes. However, simple deterministic forcing contracts will not be desired by owners when they face nontrivial uncertainty since each owner will want his manager to react to the eventual environment. Therefore, uncertainty is necessary to make the use of linear contracts in profits and sales reasonable and superior to contracts which yield the usual oligopoly outcomes.

In their paper, Fershtman and Judd (1987) show that strategic delegation is beneficial if the manager is better informed about the intercept (i.e., the consumers' maximum willingness to pay) or the slope of the demand function. In other words, it is crucial to keep in mind that the manager's information advantage is assumed. If the agent and owner ex post would have the same information, the owner could simply force the manager to implement a certain action. This, in turn, would make the use of linear contracts unnecessary. Hence, only if there are non-trivial levels of uncertainty would the owner want the manager to react to contingencies that the owner cannot observe. Nevertheless, since the agent is assumed to be risk-neutral, there are no agency costs associated with delegating decision rights to the manager.

This raises the question of whether the effects of delegating might even be beneficial if agency considerations due to moral hazard, limited liability, or adverse selection have real effects. The answer is that the firm has to carefully weigh the strategic benefits of delegation to influence competition against the agency costs emerging in the relationship with the manager. Gal-Or (1997) provides an insightful review of (early) contributions that study this trade-off. Related to the particular strategic delegation model discussed above, Fershtman and Judd (1990) demonstrate that many of the results obtained in their earlier analysis still hold if moral hazard is an issue. Merzoni (2000) shows that strategic delegation increases managerial effort and, therefore, decreases agency costs. More recently, Plehn-Dujowich and Serfes (2010) study price- or quantity-setting duopolies with public contracts where managerial effort reduces marginal costs of production. They present three interesting findings that are due to the strategic dependence (strategic complements or strategic substitutes) of firms' managerial incentives. First, they demonstrate that under strategic interaction in the product market, greater risk may actually lead a firm to provide more high-powered incentives to its manager if managerial incentives are strategic substitutes. Second, a firm's compensation scheme might be adjusted if the rival firm's idiosyncratic risk changes. Finally, if all firms are affected by a (small) decrease in risk, then managerial compensation in the industry can

change substantially if compensation schemes are strategic complements. Consequently, we might conclude that the strategic use of contracts in oligopolistic markets are an important determinant of firm performance even in the presence of agency costs.

2.2.4 Linear contracts and performance measures

Reitman (1993) continues the research of Fershtman and Judd (1987) and Sklivas (1987) by considering contracts based on profits and sales revenues. However, he shows that overly aggressive behavior and the resulting decrease in profits can be limited by incorporating stock options in the manager's compensation package (see also Spagnolo, 2000). Assuming that changes in the stock price depend on profits, the manager's compensation with stock options can be expressed as

$$w_i(q_1, q_2) = A_i + B_i[\alpha_i(\pi_i - T_i)^+ + (1 - \alpha_i)R_i].$$

This implies that a manager receives a compensation based on stock options only if profits are above a certain strike value T_i determined by the firm. As the behavior of the firms in the market tends to be more aggressive, profits decrease. The value of the stock option reduces too, possibly to the point in which the manager is not influenced any more by the value of the stock option and thus focuses on sales maximization. Consequently, for some value of the quantity chosen by the rival, each firm has a discontinuity in the best reply. The presence of the discontinuity and the threat of sales maximization equilibria induces a less aggressive behavior in the market. As Reitman points out, the result is not necessarily limited to the use of stock options. Any other form of compensation that would generate a similar discontinuity in the firms' best replies (for example, bonuses awarded only if profits reach a certain predefined level) would produce the same effects.

There is an increasing trend to use sales revenue as a performance metric in corporate executive contracts. This is documented, for example, by Huang, Marquardt, and Zhang (2015) using data of performance measures employed in executive bonus contracts for S&P 500 firms. Nevertheless, a number of papers have studied the impact of strategic delegation if managerial compensation contracts are not based on profits and sales revenues, but on different performance measures. Vickers (1985) considers a Cournot oligopoly where the owners compensate managers based on a combination of profit and sales quantities, i.e., the performance measure is $\pi_i + \alpha_i q_i$, where α_i is the incentive parameter optimally selected by the owner. Vickers' analysis confirms that in equilibrium output is higher and market price and firm profits are lower. This is in line with the intuition that we have provided above and it is based on the strategic substitutability of quantities in a Cournot game. Vickers also briefly discusses relative performance evaluation where the performance measure for manager i depends on firm i 's profit relative to the profits of all other firms. In agency settings with moral hazard, relative performance evaluation saves agency costs since common noise (which affects all agents' performance measures) can be filtered out. Salas Fumás (1992) takes a closer look at the interaction between relative performance evaluation and strategic effects of delegation and concludes that under quantity competition net profits may be higher or lower than without relative performance evaluation depending on the importance of the two effects. Under price competition, the conclusions from an agency perspective and a strategic delegation perspective differ. Filtering out common noise requires a negative weight on the rival firms' profits whereas strategic concerns would suggest a positive weight to use the contract as a

collusion device in order to keep prices near the monopoly level (see also Aggarwal and Samwick, 1999). This work is continued by Asseburg and Hofmann (2010), who investigate the strategic value of contracts (contract externalities) under relative performance evaluation in a quantity-setting duopoly. They find, for example, that the relation between the weight placed on the rival's profit and the intensity of competition (measured by the products' substitutability) is non-monotonic. Miller and Pazgal (2001) consider risk-neutral managers and show that in a differentiated-products oligopoly the following equivalence result holds: if managers' compensation contracts are based on a linear combination of own profit and rival profits, then price and quantity competition lead to the same outcomes. Their result demonstrates that in situations where the owners have sufficient control (via the contracts) over the actions chosen by their managers, the same behavior of the manager is induced, leading to the same outcomes independent of the mode of competition. If the commitment power of the owners is further increased so that a manager's strategy can be conditioned on the rivals' selected compensation schemes, then every Pareto-optimal outcome of the game can become the unique subgame-perfect Nash equilibrium of the game of strategic delegation. Consequently, for a more general class of contracts, some type of folk theorem can be shown; see Fershtman, Judd, and Kalai (1991) and Katz (2006).

Jansen, Van Lier, and Van Witteloostuijn (2007) and Ritz (2008) are the first to look at strategic delegation when contracts are based on market share. They demonstrate that (due to the relative performance component), under market share-based contracts competition is less intense than under revenue- or quantity-based contracts and, hence, profits are higher. For a more general approach, see Berr (2011). Kopel and Lambertini (2013) use linear demands discussed by Bowley (1924) and Singh and Vives (1984) and confirm that using contracts based on market share indeed makes competition softer and leads to higher profits than under revenue-based contracts. Jansen, Van Lier, and Van Witteloostuijn (2009) consider a quantity-setting duopoly and compare firm performance under profit-based, sales-based, market share-based, and relative profits-based compensation contracts. They show that compensation contracts based on relative profits lead to the highest profits of the competitors. In a recent paper, Dockner and Löffler (2015) consider a dynamic, infinite-horizon, quantity-setting duopoly. Managers are compensated through a contract that is based on a weighted sum of the discounted streams of profits and sales. Product market competition is characterized by managers choosing Markovian decision rules. The authors show that in contrast to a static setting, in a dynamic environment equilibrium incentives of the compensation contracts correspond to rivalry restraint. That is, managers are given incentives to show less aggressive behavior, resulting in lower outputs and higher profits for all firms. The reason for this contrasting result is that product market competition in Markov strategies leads to substantially higher (compared to static) output levels. The owners optimally take this aggressive behavior into account and adjust the contract accordingly.

2.2.5 Observability of contracts and strategic commitment effects

The results on strategic delegation are derived under the assumption that contracts are publicly observable. Arguably, in today's corporate governance environment the assumption that managerial compensation plans are public is not unrealistic. The adoption of managerial incentive plans commonly requires shareholder approval. Moreover, recent changes in corporate governance rules are intended to increase the transparency with regard to managerial compensation (see, e.g., Bagnoli and Watts, 2015, Larcker and Tayan, 2011). The literature has quite

extensively discussed under which conditions unobservable (private) contracts can serve as precommitments. Katz (1991) finds that unobservable contracts lose their commitment value if owner and manager have the same information *ex ante* (i.e., before signing the compensation contract), owner and manager are risk-neutral, they have the same disutility of effort, and residual claimant contracts are feasible. In other words, if it is common knowledge that a contract can solve the agency problem and owner and manager have the same preferences and capabilities, then a residual claimant contract achieves perfect delegation and the agent would behave the same way as the principal. What this neutrality result states is that if contracts are unobservable, then there is no effect on the play in the second period and the strategic value of delegation as described by Fershtman and Judd (1987) and Sklivas (1987) is lost. However, considering a wholesaler–retailer relationship (which can be interpreted in terms of managerial strategic delegation), Pagnozzi and Piccolo (2012) show that the neutrality result holds only under a specific assumption about the delegates’ conjectures on their rivals’ contracts. Under so-called passive beliefs (or passive conjectures), if manager i is offered a contract specifying a weight α_i different from the weight the manager expects in equilibrium (i.e., an out-of-equilibrium offer), manager i would not update his or her belief about the contract offered to the rival manager j . Manager i would assume that manager j ’s (unobserved) choice remains the same. Under passive beliefs, therefore, each managerial firm acts as if decisions are taken by the owner.

The limitations of contracts as a commitment device were further studied in a number of papers. For example, Fershtman and Kalai (1997) introduce incentive delegation, where a manager is given an incentive scheme and subsequently chooses the payoff-maximizing strategy (in contrast to instructive delegation studied by Katz, 1991 where the manager receives a set of binding instructions). They show that incentive delegation, even if unobservable, has commitment value. Consequently, if the manager’s contract regulates the compensation based on the game’s outcome but does not instruct the manager which strategy to choose, then delegation has strategic value even if the exact terms of the contract are unobservable to rival firms. Fershtman and Kalai further show that if there is a small probability that contracts become observable or if the strategic delegation game is repeated, commitment value is established. Corts and Neher (2003) also find that vertical delegation can have strategic value even if contracts are unobservable, if (and only if) decision rights are granted to multiple agent firms and they are given an ownership stake. What is critical here is that the choice of the organizational form (i.e., the number of agent firms and the level of ownership stake) is observable by all players. Delegation yields a strategic advantage with regard to a competing integrated structure if delegation leads to less (more) aggressive behavior and the inter-structure game is one of strategic complements (strategic substitutes). Koçkesen and Ok (2004) study sequential equilibria of one-sided delegation games in which only one owner has the option to delegate. They conclude as well that even if contracts are unobservable, delegation may arise solely due to strategic reasons in a general class of economic settings using well-supported equilibrium as a refinement. Koçkesen (2007) provides an extension to two-sided delegation games where both owners have the option to delegate. Finally, Bagwell (1995) analyzes a leader–follower model where the follower can only imperfectly observe the choice of the leader. He shows that even with the slightest degree of uncertainty, the value of commitment is lost. Maggi (1999), however, shows that value of commitment is restored if the leader has private information, e.g., about a cost parameter only relevant to the leader. The reason is that the follower uses the signal about the leader’s action to infer the leader’s type

and the quantity that has been chosen. Hence, the leader has an incentive to produce more than the Cournot output to influence the follower's belief about the leader's type.

If contracts are private, competition among oligopolists on the product market and internal governance might still interact in important ways. For example, Chalioti (2015) studies R&D investments under quantity competition in a duopoly. Firms design incentive contracts for their managers based on realized cost reductions. Managers choose unobservable efforts to reduce production costs. Finally, firms determine production levels. The author shows that due to the agency problem between the firm and its risk-averse manager, the incentive for reducing production costs is dampened. This might yield higher profits than in a situation without moral hazard. In Bhardwaj (2001), demand depends on prices and on efforts put in by (risk-averse) sales representatives. The decision of firms to delegate the price choice is observable by the rival firm. Although the contract between the firm and the sales representative is not observable for rival firms, price delegation serves as a commitment to soften competition if price competition is intense. In an incomplete contracting framework, De Bettignies (2006) demonstrates that the choice of an (un)integrated structure interacts with the degree of product market rivalry. Another stream of literature is concerned with the question of how agency issues within a firm are affected by a change in competition in the product market. This literature, for example, addresses the question of whether managers in firms expend more or less effort if the competitiveness of the industry (measured, e.g., by the number of firms or the substitutability between products) changes. The focus here is on the relation between managerial and organizational slack and the structure of the industry or the mode of competition. See, e.g., Piccolo, D'Amato, and Martina (2008), Raith (2003), Baggs and De Bettignies (2007), Graziano and Parigi (1998), and Schmidt (1997).

2.2.6 Take-it-or-leave-it contracts or owner-manager bargaining

In contrast to a take-it-or-leave-it contract, Fershtman (1985) studies a duopolistic environment and assumes that in each firm i there are two managers, where one tries to maximize profits and the other tries to maximize sales quantity. Owners want to maximize profits. They can organize their firm i such that only the profit-maximizing manager determines the output level (imagine that an incentive scheme based on profits is written in this case) or that the two managers determine the output level by Nash bargaining. Fershtman shows that (under identical costs, $c_1 = c_2$, and no hiring costs, $U = 0$) both owners prefer the bargaining solution. This, however, should not come as a surprise since bargaining between the two managers (where one is a profit-maximizer) is identical to bargaining between a profit-maximizing owner and a sales-maximizing manager. The outcome with (efficient) bargaining between owner and manager is identical to the outcome with a take-it-or-leave-it contract in Fershtman and Judd (1987). More recently, the issue of bargaining between each owner-manager pair under oligopolistic competition has gained increasing interest. For example, Balasubramanian and Bhardwaj (2004) demonstrate that the conflicting objectives of manufacturing and marketing managers can actually result in higher profits than under full coordination. In line with Fershtman (1985), they assume that the firm's owner designs incentive contracts for the firm's managers and that the managers (Nash) bargain to obtain a compromise solution for the price and the quality of the product. Kopel, Pezzini, and Rezzi (2016) discuss further contributions on bargaining in agency and delegation models. They demonstrate that in a location-price Hotelling-type model outcomes and payoffs are unaffected if owners and managers bargain over the full terms of the incentive contracts,

i.e., the salary component A_i , the bonus B_i , and the incentive rate α_i and they compare their findings to the contrasting results of Van Witteloostuijn, Jansen, and Van Lier (2007).

2.2.7 Empirical and experimental evidence on strategic delegation

There is considerable empirical evidence that industry characteristics play an important role in the design of compensation contracts for managers and for managerial incentives in general (see, e.g., Karuna, 2007, Cuñat and Guadalupe, 2005). In particular, Kedia (2006) finds support for the qualitative predictions of the strategic delegation approach. Using data on profits and sales, he shows that if the firms' choices are strategic substitutes (i.e., the firms' best replies are negatively sloped), then the pay-for-performance incentives for the firms' CEOs are significantly lower. That is, in a CEO's incentive contract more weight is put on sales and less on profits. If the firms' choices are strategic complements (i.e., the firms' best replies are positively sloped), then the pay-for-performance incentives for the firms' CEOs are significantly higher. Irwin (1991) finds further evidence for the use of compensation contracts as a device to achieve a competitive advantage. He presents details about the seventeenth-century Anglo-Dutch rivalry between the English East India Company and the Dutch East India Company. His empirical analysis confirms the hypothesis that the Dutch government used a monopoly charter (as a commitment) to provide managerial incentives to achieve a Stackelberg leader position. On a broader level, some papers investigate the strategic choice of organization design under oligopolistic competition. For example, Vroom and Gimeno (2007) provide evidence that multi-unit organizations employ different ownership forms (company ownership versus franchising) to commit to more or less aggressive competitive behavior. Using data from the Texas hotel industry, they show that company-owned units (due to their more rigid rules) leads to firms committing to less aggressive pricing behavior and soften competition in concentrated markets. Sengul and Gimeno (2013) address the problem of negative intra-firm spillovers if subsidiaries of multi-industry firms compete in the same multiple markets. They argue that in order to engage subsidiaries and at the same time restrict the effect of these negative spillovers, firms use constrained delegation, i.e., firms delegate business-level decisions while constraining the subsidiaries' decision rights and available resources. Their empirical analysis supports their arguments. Finally, Slade (1998) finds empirical evidence of strategic motives for vertical separation in the retail gasoline market.

There is little, but some experimental work on the strategic effects of delegation. Fershtman and Gneezy (2001) study the use of delegates in an ultimatum game. They find that the proposer's payoffs are significantly higher when a delegate makes the offer to the responder as it induces a change in the responder's behavior. Huck, Müller, and Normann (2004) more directly test the predictions of the strategic delegation models that firms use properly designed incentive contracts to make their managers more aggressive. Their experiment shows that owners rarely choose the contract predicted by the delegation model (note that this is ex post rational as profits are higher than in the case where both firms delegate). The authors point out that collusion between owners can partially explain the results. However, the main driver seems to be that managers are more (less) aggressive in strategically weak (strong) positions than theory predicts. Hence owners are less inclined to provide managerial incentives for aggressive behavior. In a more recent paper, Du, Heywood, and Ye (2013) carry out an experiment using a mixed duopoly market where a public (welfare-maximizing) firm competes against a profit-maximizing firm. Since welfare is higher if outcomes are more competitive, in contrast to a Cournot duopoly studied by Huck et al. (2004) firms do not

have an incentive to collude. The laboratory results broadly confirm the predictions of the delegation model: both the private and the public firms adopt the weights of the equilibrium delegation contract. Finally, Barreda-Tarazona et al. (2016) consider the quantity-setting duopoly with private firms under strategic delegation. In accordance with the theoretical results, they find that subjects choose relative performance contracts more frequently than contracts based on profits and sales revenues. Furthermore, their experimental evidence confirms that firms' owners indeed set incentives for their managers to induce more aggressive market behavior.

3 DELEGATION AND VERTICAL SEPARATION

Our main focus so far has been the description of the strategic effects that the delegation of decision making to a manager may generate and its impacts on the payoffs of the owner of a firm. Clearly the same reasoning can be extended to other aspects of firms' conduct and organization. If a manufacturing firm decided to be vertically separated and to allow a retailer to independently set the price of the final goods sold to consumers, then the pricing decisions of the rivals will be affected by the actions of the retailer. The decision to use an independent retailer can indeed be classified as a form of strategic delegation of decision making. The contractual relationship in this case would be based on wholesale prices and (in some cases) franchise fees that will affect the strategic behavior of the retailers in the same way as managerial incentive contracts affect managerial decisions. The intuition is that a manufacturer can choose the wholesale price with the intention to strategically turn the (exclusive) retailer into a more or less aggressive competitor in the market. If the contract between the manufacturer and the retailer is publicly known and renegotiations are not possible, then vertical separation works as a credible commitment device in oligopolistic competition.

3.1 Vertical Separation and Stackelberg Price Leadership

Bonanno and Vickers (1988) highlight the strategic commitment effects in a model of vertical separation. In line with their work, the present section concentrates on the optimal design of the distribution channels and the optimal choice of the contract that governs the relationship between a manufacturer and a retailer if the final product market is imperfectly competitive. To illustrate, we consider a standard Hotelling model. The product market is characterized by a linear city where consumers are uniformly distributed along the interval $[0, 1]$ and two single-product firms, say firm 1 and firm 2, manufacture differentiated products. The two manufacturers are located at the extremes of the unit interval; without loss of generality suppose that firm 1 is located at $x_1 = 0$ and firm 2 is located at point $x_2 = 1$. Each firm has the option to be the retailer of its own product (vertical integration) or to sell through an independent retailer (vertical separation). In the former case, the manufacturer offers the product on the market and sets its product's price, p_i . In the case of vertical separation, the manufacturer i determines the contract terms at which it will supply its (exclusive) retailer. The two-part tariff consists of a franchise fee, T_i , plus a wholesale price per unit, w_i . The retailer then sets the market price for the product to maximize own profits. The transportation cost paid by a consumer who is located at $x \in [0, 1]$ and buys firm 1's (firm 2's) product is tx^2 ($t(1 - x)^2$).

Assuming that each consumer buys exactly one product, the utility of a particular consumer located at $x \in [0, 1]$ is

$$u_x = \begin{cases} v - tx^2 - p_1 & \text{if consumer buys from firm 1,} \\ v - t(1-x)^2 - p_2 & \text{if consumer buys from firm 2,} \end{cases}$$

where $v > 0$, the gross benefit for the consumers to purchase a unit of good, is sufficiently high to ensure full market coverage. Note that an increase in transportation costs, i.e., a higher value of t , can be interpreted as capturing a lower degree of competition. The marginal consumer \hat{x} who is indifferent between buying from firm 1 and buying from firm 2 is given by

$$\hat{x}(p_1, p_2) = \frac{1}{2} + \frac{p_2 - p_1}{2t}.$$

Therefore, in an interior solution, the demand for firm i 's product is

$$\begin{aligned} q_1(p_1, p_2) &= \hat{x}(p_1, p_2), \\ q_2(p_1, p_2) &= 1 - \hat{x}(p_1, p_2). \end{aligned}$$

In the case that manufacturer i stays vertically integrated (superscript I), its profit is given by

$$\pi_i^I(p_1, p_2) = (p_i - c_i)q_i,$$

where c_i denotes the marginal production cost. (Again we assume that marginal costs c_i are such that quantities derived from equilibrium prices are nonnegative.) In the case that manufacturer i decides to get vertically separated (superscript S), the payoffs of the manufacturer and its retailer (or franchisee) are, respectively, given by

$$\begin{aligned} \pi_i^S &= T_i + w_i q_i - c_i q_i, \\ F_i &= (p_i - w_i) q_i - T_i. \end{aligned}$$

As in the previous section, we explicitly include the case where only one firm vertically separates. The timing of this game is as follows. First, manufacturers simultaneously choose their organizational structure, i.e., to vertically separate ($m_i = 1$) or not ($m_i = 0$). In case of vertical integration, manufacturer i selects the profit-maximizing price p_i . In the case of vertical separation, manufacturer i chooses the optimal contract parameters, i.e., the franchise fee T_i and the wholesale price w_i such that π_i^S is maximized subject to the participation constraint of the retailer, $F_i \geq U$. Obviously, the franchise fee T_i will be adjusted so that in equilibrium the participation constraint binds and hence, $T_i = (p_i - w_i)q_i - U$. As a consequence, the manufacturer maximizes the net profit $\pi_i^S = (p_i - c_i)q_i - U$. The wholesale price w_i is used to provide (strategic) incentives to the retailer. Finally, retailer i maximizes its profit F_i with respect to the price p_i .

At this point it should become obvious that the structures of the strategic delegation game with a manager and the vertical separation game presented in this section are identical. In fact, alternatively one can think of the decision of a firm to hire a manager who is then in charge of setting the market price of a differentiated product. To provide (strategic) incentives, the firm

Table 10.2 Normal-form game to determine the overall equilibrium of the game

		Manufacturer 2	
		$m_2 = 0$	$m_2 = 1$
Manufacturer 1	$m_1 = 0$	π_1^I, π_2^I	$\pi_1^F, \pi_2^L - U$
	$m_1 = 1$	$\pi_1^L - U, \pi_2^F$	$\pi_1^S - U, \pi_2^S - U$

offers a compensation contract to its manager. The outcomes and payoffs of this alternative game with public contracts would be identical.

The four subgames for all possible choices in period 1 can be easily solved by backward induction. To determine the overall outcome of the game, the normal-form game depicted in Table 10.2 is used. The entries in the cells of the payoff matrix correspond to the manufacturer's payoffs in each of the four subgames. Again, we provide brief comments on the entries below.

Both firms stay vertically integrated If both firms stay vertically integrated ($m_1 = m_2 = 0$), then the game corresponds to a standard Hotelling game with heterogeneous marginal costs. Prices, market shares and profits in equilibrium are $p_i^I = (3t + 2c_i + c_j)/3$, $q_i^I = (3t - c_i + c_j)/6t$, $\pi_i^I = (3t - c_i + c_j)^2/18t$.

Only one firm separates Assume that without loss of generality only firm 1 separates and delegates the price choice to an exclusive retailer, i.e., $m_1 = 1, m_2 = 0$. In the final stage of the game, firm 1's retailer chooses the price such that F_1 is maximized. Depending on the wholesale price w_1 being smaller or larger than the marginal production cost c_1 , the retailer chooses a higher or lower price than the firm would under centralized decision-making. Manufacturer 2 is selecting the price so that its profit is maximized. The resulting price best replies are

$$p_1 = \frac{t + w_1}{2} + \frac{1}{2}p_2$$

$$p_2 = \frac{t + c_2}{2} + \frac{1}{2}p_1.$$

Solving yields similar prices as in the integrated case,

$$p_1(w_1) = \frac{3t + 2w_1 + c_2}{3}$$

$$p_2(w_1) = \frac{3t + 2c_2 + w_1}{3},$$

where c_1 is replaced by the wholesale price w_1 .

The reduced-form profits can be written as $\pi_i(p_1(w_1), p_2(w_1))$. Again, one could wonder why manufacturer 1 would want to set a retail price w_1 not equal to marginal production cost c_1 since this would lead to identical outcomes as in the case of vertical integration. It turns out that by setting $w_1 > c_1$ both firms would select higher prices compared to the integrated

outcome. Differentiating the reduced-form profits, using the envelope theorem, and evaluating at $w_1 = c_1$ yields

$$\underbrace{\frac{\partial \pi_1}{\partial p_1} \frac{\partial p_1}{\partial w_1}}_{=0} + \underbrace{\frac{\partial \pi_1}{\partial p_2} \frac{\partial p_2}{\partial w_1}}_{>0 >0} > 0.$$

Likewise,

$$\underbrace{\frac{\partial \pi_2}{\partial p_1} \frac{\partial p_1}{\partial w_1}}_{>0 >0} + \underbrace{\frac{\partial \pi_2}{\partial p_2} \frac{\partial p_2}{\partial w_1}}_{=0}.$$

The result clearly resembles the strategic effects of strategic delegation under price competition, as we have discussed in the previous section. This shows that an increase in the wholesale price increases the profits of both firms, $\left. \frac{d\pi_i}{dw_1} \right|_{w_1=c_1} > 0$. The best reply of both firms slope upwards, i.e., prices are strategic substitutes. Increasing w_1 shifts firm 1's best reply upwards, but leaves the price best reply of firm 2 unchanged (cf. Figure 10.2, point E_2). Consequently, by selecting the wholesale price appropriately, the manufacturer can induce the retailer to select the same market price as a Stackelberg price leader. Manufacturer 1 uses its retailer as a device to collude with its rival and to obtain higher profits. Recall that under quantity competition the firm designs the contract to induce a more aggressive behavior from its delegate. This corresponds to the manufacturer choosing a wholesale price $w_1 < c_1$. As price undercutting induces the rival to lower its price as well, under price competition top dog behavior is not a profitable strategy. Instead the manufacturer's wholesale price is higher than marginal cost, which induces the retailer to raise its market price. Although this "double markup" is problematic in a single firm environment, under strategic interaction this "soft" behavior of the retailer causes the rival firm to raise its price as well. Vertical separation is an example of a so-called "puppy dog" strategy (see again, Fudenberg and Tirole, 1984).

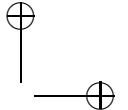
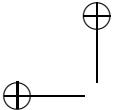
Solving the first-order condition, $d\pi_1/dw_1 = 0$, yields

$$w_1 = \frac{3t + 3c_1 + c_2}{4},$$

and the resulting market shares, prices, and profits are

$$\begin{aligned} q_1^L &= \frac{3t - c_1 + c_2}{8t}, & q_2^F &= \frac{5t - c_2 + c_1}{8t}, \\ p_1^L &= \frac{3t + c_1 + c_2}{2}, & p_2^F &= \frac{5t + 3c_2 + c_1}{4}, \\ \pi_1^L &= \frac{(3t - c_1 + c_2)^2}{16t}, & \pi_2^F &= \frac{(5t - c_2 + c_1)^2}{32t}. \end{aligned}$$

Note that $w_1 > c_1$ if and only if $q_1^L > 0$. Observe that under price competition, firm 2 may enjoy a second-mover advantage. This is apparent in the case of identical marginal costs,



$c_1 = c_2 = c$. The firm that delegates via vertical separation tends to select a relatively high price and this induces (due to strategic complementarity) the rival firm to increase its price. However, the price of the vertically integrated firm may be lower than the rival's price (indeed $p_1^L > p_2^F$ if firms have identical costs) and this allows the integrated firm to enjoy an expansion of sales. As a consequence, this may lead to $\pi_2^F > \pi_1^L$.

Both firms separate The subgame where both firms separate can be solved in an analogous way. The best replies of the firms' retailers are $p_i = (t + w_i + p_j)/2$ and both manufacturers try to achieve Stackelberg price leadership by increasing their wholesale prices,

$$w_i^S = \frac{5t + 4c_i + c_j}{5}.$$

As a result, competition is softened and market prices are increased,

$$p_i^S = \frac{10t + 3c_i + 2c_j}{5}.$$

This yields the following profits for the manufacturers,

$$\pi_i^S = \frac{(5t - c_i + c_j)^2}{25t}.$$

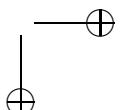
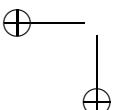
Having completed the entries of the payoff matrix in Table 10.2, we can now check which organizational structure is most beneficial for the firms under which conditions. First note that for $U < \min[\pi_1^S - \pi_1^F, \pi_2^S - \pi_2^F] = \min[\frac{7(5t - c_1 + c_2)^2}{800t}, \frac{7(5t - c_2 + c_1)^2}{800t}]$ vertical separation is an equilibrium. If, in addition, $U < \min[\pi_1^L - \pi_1^I, \pi_2^L - \pi_2^I] = \min[\frac{(3t - c_1 + c_2)^2}{144t}, \frac{(3t - c_2 + c_1)^2}{144t}]$ holds, then separation is a dominant strategy for both firms. If, on the other hand, $\frac{7(5t - c_2 + c_1)^2}{800t} < U < \frac{(3t - c_1 + c_2)^2}{144t}$, then firm 1 vertically separates while firm 2 optimally stays integrated. Under these conditions, vertical separation, that is, delegating the choice of the market price to a retailer and selecting an appropriate franchise fee and wholesale price, endogenously gives firm 1 the position of a Stackelberg price leader.

3.2 Assumptions of the Model and Extensions

As should be clear by now, the model with vertical separation described above embodies the standard features of strategic delegation as described in Section 2 and it has the same restrictions and limitations as the strategic delegation model with a manager. Therefore, we shall only briefly touch on the key assumptions and possible extensions of the model and report on noteworthy related contributions in the literature.

3.2.1 Linear wholesale prices and alternative vertical restraints

We have assumed that an upstream firm offers its downstream manufacturer a non-linear (or two-part) tariff. The contract stipulates a franchise fee and a per-unit wholesale price. Alternatively, other contract types are possible. For example, under linear pricing the manufacturer offers the retailer a per-unit wholesale price without a franchise fee. Under this



contract, the manufacturer cannot use the fixed fee to extract all the profit from the retailer. The questions then arise whether the manufacturer's profit would be higher and which contract type would be chosen in equilibrium given that both retailers have a choice.

Consider the situation where both manufacturers choose linear pricing. At the final stage of the game, the retailers would select identical price reactions as above, i.e., $p_i = (t + w_i + p_j)/2$ to maximize $F_i = (p_i - w_i)q_i$. The constant franchise fee does not have any influence at this stage. However, under linear pricing retailer i would select w_i such that $\pi_i^S(w_1, w_2) = (w_i - c_i)q_i$ is maximized. In the case where $U = 0, c_1 = c_2 = c$, this would result in the following profits. If both firms stay vertically integrated, then profits are $\pi^I = t/2$. If they separate and use two-part tariffs, then profits are t . In contrast, if firms separate and use linear pricing, then profits are higher, $\pi^S = 3t/2$. This analysis suggests that the more flexible two-part tariff has less commitment value than a linear pricing contract. It also demonstrates that the effect of double marginalization is overcome by the effect of softening competition by delegating the price choice to independent retailers. If the contract type is itself an endogenous choice of the manufacturers, then Rey and Stiglitz (1995) show that choosing a contract with a franchise fee dominates a linear pricing contract. The manufacturers are trapped in a prisoner's dilemma, however. That is, both firms offer their retailer two-part tariffs although both firms would earn higher profits if they could coordinate to use linear pricing instead. Like non-linear tariffs, other vertical restraints like (publicly observable) exclusive territorial agreements can serve to dampen competition and can facilitate collusion even if interactions are not repeated. For example, with exclusive territories the retailers are granted the right to exclusively sell its product in a specified region or to a group of customers. Granting exclusive territories eliminates intra-brand competition and is an effective commitment device to reduce inter-brand competition and hence to increase manufacturers' profits even without franchise fees (Rey and Stiglitz, 1988, 1995). In comparison with non-linear contracts, it is more plausible that manufacturers can observe whether rivals have assigned exclusive territories to their retailers (Rey and Vergé, 2008). For further details, see also Irmén (1998).

3.2.2 Observability of contracts and mode of competition

Since manufacturers can obtain higher profits with vertical separation and observable two-part tariff contracts that govern the relationship with their retailer (compared to the integrated solution), seemingly manufacturers have an incentive to disclose this information. On the other hand, if contracts were unobservable in the vertical separation setting presented above, if there is no (ex ante) asymmetry of information between the retailer and the manufacturer, and if two-part tariffs would be possible, then the wholesale prices would be identical to marginal production costs (again the neutrality result that we have described in the previous section holds). Consequently, the outcome would be the same as under vertical integration. With identical marginal production costs, $c_1 = c_2 = c$ and $U = 0$ the outcome with private contracts (and passive conjectures) would be therefore $p_i = t + c, q_i = 1/2, \pi_i = t/2$. Recall that this neutrality result holds only under passive beliefs where retailer i does not revise its belief upon observing a contract offer that is different from what the retailer has expected in equilibrium.

Pagnozzi and Piccolo (2012) argue, however, that in certain situations with private contracts it is more realistic to assume symmetric beliefs. Under symmetric beliefs, retailers may conjecture (because of, for example, some form of bounded rationality) that identical

wholesalers always offer the same retail contract. In other words, if retailer i receives a contract offer that differs from what this retailer expects in equilibrium (an out-of-equilibrium offer), then this retailer believes that the other retailer j receives the same offer. If retailers conjecture that other retailers are offered the same contract, then delegation creates a “belief effect”: the wholesale price offered by the manufacturer affects the retailer’s conjecture about the behavior of the rival. As a consequence, under symmetric (or partially symmetric) belief separation can have commitment effects, even when contracts are not observable. In our Hotelling model with vertical separation of both firms, assuming $c_1 = c_2 = c$ and $U = 0$ the argument unfolds as follows. Retailer i ’s best reply is as before, $p_i = (t + w_i + p_j)/2$, but retailer i ’s conjecture about retailer j ’s price would be $\hat{p}_j = (t + w_i + p_i)/2$ (since retailer i believes that retailer j receives the same contract offer). This leads to the price $p_i = t + w_i$ and likewise $\hat{p}_j = t + w_i$. Of course, in a symmetric equilibrium the managers’ beliefs must be consistent with strategies, i.e., $\hat{p}_k = p_k = \hat{p}$. Manufacturer i now maximizes $\pi_i^S = T_i + (w_i - c)q_i(\hat{p}(w_i), \hat{p}(w^*))$ with regard to the wholesale price w_i subject to the retailer’s participation constraint $F_i = (\hat{p}(w_i) - w_i)q_i(\hat{p}(w_i), \hat{p}(w_i)) - T_i \geq 0$. Note that the manufacturer takes retailer j ’s price as fixed, since the manufacturer – in equilibrium – expects the rival manufacturer to offer w^* and its retailer to respond with $\hat{p}(w^*)$. In contrast, retailer i ’s belief $\hat{p}(w_i)$ depends on the retail price w_i since this retailer believes that the other retailer gets the same offer. In equilibrium, $F_i = (t + w_i - w_i)\frac{1}{2} - T_i = 0$, so it follows that $T_i = t/2$. Consequently, manufacturer i maximizes $\pi_i^S = \frac{t}{2} + (w_i - c)[\frac{1}{2} - \frac{1}{2t}(t + w_i - t - w^*)]$ and the first-order condition yields the equilibrium wholesale prices $w^* = t + c$. The resulting retail prices and firms’ profits are $p^* = \hat{p}(w^*) = 2t + c > t + c$ and $\pi^* = t > t/2$. It is important here that under symmetric beliefs each manufacturer – by increasing the wholesale price – can use its retailer’s belief to increase the market price. As a consequence of this “belief effect”, each manufacturer’s profit is higher than under centralization (or separation under passive beliefs).

Under price competition, vertical separation serves to commit to higher retail prices by charging the retailers wholesale prices above marginal costs (and franchise fees are used by the manufacturers to extract the higher profits). In contrast, under quantity competition each manufacturer has an incentive to make its retailer more aggressive. The situation is identical to owners who delegate quantity decisions to their managers and offer them publicly observable compensation contracts with some positive weight on sales revenues. This induces managers to act as if marginal production costs are smaller than the actual marginal costs c . To illustrate the effect of vertical separation of both firms, consider a simple homogeneous product market as in the managerial delegation model in Section 2. Let $c_1 = c_2 = c$ and $U = 0$. Each manufacturer would maximize $F_i = (p - w_i)q_i - T_i$, which leads to $q_i = (a - 2w_i + w_j)/3b$. Taking into consideration that the participation constraint of a retailer binds in equilibrium, the manufacturer maximizes $\pi_i^S = (p(q_i, q_j) - c)q_i$, which leads to $w_i^e = (6c - a)/5 < c$. The resulting profits in equilibrium are $\pi^e = 2(a - c)^2/25b$, which coincides with our earlier findings and shows that vertical separation under quantity competition leads to lower profits than vertical integration.

Recall that with private contracts and passive beliefs, the neutrality result yields that profits under separation and integration coincide since contracts do not have strategic effects. However, if retailers hold symmetric beliefs, the “belief effect” can lead to the surprising result that under downstream quantity competition wholesalers can achieve full collusion in equilibrium. The intuition is that increasing the wholesale price w_i leads retailer i to

decrease quantity q_i (since marginal retail costs are increasing) and, at the same time, this induces retailer i to believe that retailer j is reducing its quantity too. The manufacturers can exploit the belief effect to induce their retailers to decrease their quantities and, in equilibrium, achieve full collusion. For a simple linear homogeneous-products duopoly with vertical separation, under symmetric beliefs retailer i selects the quantity $q_i(w_i)$ assuming that manufacturer j offered its retailer the same wholesale price. Hence, $q_i(w_i) = \frac{a-w_i}{2b} - \frac{1}{2}\hat{q}_j$, where $\hat{q}_j = \frac{a-w_1}{2b} - \frac{1}{2}q_i$. This yields $\hat{q}(w_i) = (a - w_i)/3b$. The manufacturer selects w_i such that its profit $w_i\hat{q}(w_i) + T_i$ is maximized, where the franchise fee (in equilibrium) follows from the (binding) retailer's participation constraint, $T_i = [a - b(\hat{q}(w_i) + \hat{q}(w_i)) - w_i]\hat{q}(w_i)$. Solving the first-order condition leads to $w^* = \frac{a+3c}{4}$. As a consequence, the resulting quantities and profits in equilibrium are $\hat{q}(w^*) = (a - c)/4b$ and $\pi^* = (a - c)^2/8b$. In other words, given that contracts are private, under symmetric beliefs the firms can achieve full collusion and share the monopoly profit. Hence, in this situation the firms' profits are higher than under centralization and under delegation with public contracts. Both owners do not have an incentive to disclose the details of the contract to their rival. For further details we refer to Pagnozzi and Piccolo (2012).

The importance of unobservable contracts and private information in models of vertical separation has been studied intensively in the literature. For example, in a recent paper Bassi, Pagnozzi, and Piccolo (2015) study a location model where the retailers choose their location and are privately informed about their costs. They show that under the assumption of passive beliefs, the principle of maximum differentiation does not hold. O'Brien and Shaffer (1992) show that with unobservable contracts under passive beliefs, the vertically integrated outcome with non-linear contracts cannot be achieved. They argue that this explains the use of vertical restraints like exclusive territories or resale price maintenance. Rey and Stiglitz (1995) show that if wholesale contracts are unobservable, then the commitment to exclusive territories without franchise fees can still lead to higher profits and be chosen by manufacturers in equilibrium. The reason is that double marginalization in this case leads to higher wholesale prices and this softens competition downstream. If the induced loss of profits due to double marginalization is not too large, then it pays off to vertically separate and assign exclusive territories to the retailers despite the fact that contracts between the manufacturer and the retailer are unobservable. Under these circumstances, delegation is beneficial but not for strategic reasons (Katz, 1991, Irmel, 1998, Caillaud and Rey, 1995). As mentioned before, if additional agency costs arise if pricing authority is delegated (e.g., since the retailer is better informed about demand conditions or retailing costs), then the firm has to trade off the benefits of delegation as a mechanism to dampen competition against the agency costs that arise in the relationship with the retailer (Gal-Or, 1997).

3.2.3 Design of distribution channels and strategic divisionalization

In the marketing and management science literature, the optimal design of the distribution channels under oligopolistic competition has been addressed, among many others, by Moorthy (1988), Coughlan and Wernerfelt (1989), Choi (1991, 1996), and Gupta and Loulou (1998). Moorthy (1988) considers strategic decentralization where each manufacturer charges its retailer a wholesale price per unit (but no franchisee fee). He finds that decentralization is profitable for a manufacturer if the final products are substitutes and retailers' prices and manufacturers' prices are strategic complements. If products are complements, then prices have to be strategic substitutes. Coughlan and Wernerfelt (1989) raise some concerns about

the implicit assumptions in the literature on distribution channel management. Similarly to Katz (1991), they show that if intra-channel contracts are unobservable to competitors, then outcomes are identical to the centralized solution. Choi (1991, 1996) studies several channel structures, e.g., a common retailer or a common manufacturer, in a price-setting duopoly. He compares outcomes and payoffs for the manufacturers and retailers under varying degrees of product substitutability. Gupta and Loulou (1998) study the design of the distribution channel under the assumption that the manufacturer can invest in process innovation to reduce production costs. They find that channel equilibria under linear wholesale contracts depend on the interplay between product differentiation and R&D efficiency (i.e., the ease of reducing marginal production costs) and provide refinements of earlier results in the literature.

Bhardwaj and Balasubramanian (2005) combine the topics of vertical separation and strategic delegation and analyze the impact of strategic incentives on equilibrium outcomes and profits and on optimal channel structure decisions. In contrast to the literature, they find that asymmetric channel structures with one manufacturer selling through a profit-maximizing retailer and the other vertically integrated manufacturer providing strategic incentives to its manager can be an equilibrium (see also Chou, 2014). More recently, Anderson and Bao (2010) study a model with n competing supply chains and provide conditions under which decentralization leads to higher profitability of manufacturer, retailer, or the whole industry. Matsui (2012) presents a contrasting result. He shows that an incumbent might want to integrate in order to deter entry. Liu and Tyagi (2011) demonstrate that (strategically) delegating production to an upstream supplier might be beneficial for retailers who are in charge of setting product positioning and prices. Further references to the literature on strategic outsourcing can be found in Kopel, Löffler, and Pfeiffer (2016).

Another important stream of literature on strategic organization design considers the incentives for creating independent divisions under oligopolistic competition. For example, Baye, Crocker, and Ju (1996) study a two-stage game where firms first select the number of independent divisions (where the setup costs are linear in the number of firms) and then divisions engage in Cournot competition. They find that with linear demand and costs, if the number of firms is larger than two, firms typically choose more divisions than is socially optimal. Similar incentives like under strategic delegation drive this result where commitment is achieved here by setting up a larger number of independent divisions. Creating independent divisions creates intra-firm competition but allows the firm to commit to a Stackelberg leader outcome. A major difference is that even with two firms the same equilibrium outcome as under perfect competition can be obtained if divisionalization costs tend to zero whereas in the strategic delegation model this only happens if the number of firms tends to infinity. Ziss (1998) studies an extension with product differentiation. González-Maestre (2000) considers a combination of the divisionalization and strategic delegation models. He assumes that firms can create independent divisions or franchisees. These divisions are then hiring managers to determine the quantities and give the managers (strategic) incentive compensation contracts. A firm's objective is to maximize the sum of the profits of its divisions. In this divisionalization-delegation game, González-Maestre (2000) finds that there is no incentive to divisionalize in a duopoly and that in a market with three firms the number of divisions is below the social optimum. Rysman (2001) addresses the point that under duopolistic quantity competition, setting up independent divisions and signing a two-part tariff contract with a single franchisee can both be used to increase output. He shows that both tools are perfectly interchangeable if a certain output should be induced. In equilibrium, however, both firms first choose to have

a single franchisee and then use non-linear contracts for inducing the desired (Stackelberg) output. These findings have some similarity with the results of Martinez-Giralt and Neven (1988) and Tabuchi (2012) for the Hotelling model.

3.2.4 Strategic transfer pricing

Firms commonly use transfer prices to coordinate the decisions of independent divisions and to align the divisions' objectives with the goal of the firm. However, similarly to wholesale prices in the vertical separation setting, transfer prices might also be used to determine strategic incentives of division managers in an oligopolistic market. Alles and Datar (1998) consider two firms that each consist of a production and a marketing division. The transfer price for the product that is transferred from the production to the marketing department is determined by the CEO of each firm (to maximize the firm's profit). Each marketing division is in charge of setting the market price of its firm's product and each division manager maximizes the profit of the division. Alles and Datar (1998) show that the equilibrium transfer prices are above marginal cost, which increases the input costs of the marketing department and induces the marketing manager to set a higher market price. Setting higher market prices leads to higher profits for both firms. The logic for this result is the same as in the vertical separation setting with independent upstream and downstream units. The firms commit to delegate the pricing decision to its marketing manager who maximizes transfer price-based divisional profits. This induces less aggressive pricing behavior by the firm's manager and its rival manager. Of course, this immediately raises the question of observability of transfer prices. Göx (2000) shows that under certain conditions firms can profitably use the choice of their cost system (absorption costing) to signal the use of transfer prices above marginal cost to their rivals. Dürr and Göx (2011) consider an international duopoly and investigate if firms can benefit from committing to the same transfer price for tax and managerial purposes (a one set of books policy) versus using separate transfer prices for the two objectives (a two sets of books policy). Intuitively, they show that a one set of books policy is profitable if competition in the market is high (if, e.g., the firms' products are less differentiated). In this case, the commitment effect of using a single transfer price outweighs the benefits of increased coordination through two distinct transfer prices. For further details about analytical research on transfer pricing, we refer the reader to Göx and Schiller (2007).

3.2.5 Take-it-or-leave-it contracts or manufacturer–retailer bargaining

Milliou and Petrakis (2007) study the impact of contract types and bargaining between the manufacturer and the retailer about the contract terms on the equilibrium upstream market structure (i.e., the incentives of upstream firms to merge). They consider quantity competition and show, for example, that upstream firms remain separated under two-part tariff contracts, but prefer to merge under wholesale price contracts. The driver of this result is as follows. When firms bargain over two-part tariffs, then the vertical chain commits to more aggressive behavior than under vertical integration. The equilibrium wholesale price is always lower than the upstream marginal cost. As a consequence, this reduces total surplus for the firms. This is intuitive and coincides with our insights of the previous section. In contrast, under wholesale price contracts, the equilibrium wholesale prices exceed the upstream marginal cost. The reason is that under wholesale price contracts, the only available instrument for transferring rents to the upstream firm is the wholesale price. Although an upstream firm's bargaining position is increased by a two-part tariff, the reduction in surplus dominates.

4 STRATEGIC DELEGATION, MANAGER TYPES, AND SUPPLY-SIDE RELATIONS

In his early paper on strategic delegation, Vickers (1985) discusses an agent appointment game. Commitment is achieved not by writing a publicly observable incentive contract for a manager or by having an independent retail division, but by hiring a particular (optimal) type of manager. Vickers considers agents whose utility is given by $\pi_i + \theta_i q_i$ where π_i and q_i represent, respectively, the profit and quantity of firm i . Under the assumption that the type of manager θ_i is observable, he derives the optimal type of manager in a linear oligopoly. He further demonstrates that a managerial firm with an optimal type of manager is more profitable than a profit-maximizing firm. Obviously, selecting the optimal type of manager is akin to determining the optimal value of the incentive parameter α_i as we have described in Section 2. In this section we first take a closer look at the effects of hiring a particular (e.g., a biased or overoptimistic) type of manager on product market competition. We then consider monopolistic downstream markets and focus on the impact of hiring a charitable manager if upstream input markets are taken into consideration. It will turn out that there are important strategic effects on the firm's supply side that are typically not considered in delegation models. This topic gives rise to a broader discussion of the optimal design of organizations if vertically related markets play a role.

4.1 Manager Types and Product Market Competition

Several other contributions have continued Vicker's line of research. Miller and Pazgal (2002) consider managers who care about relative profits where the weight on the other firm's profit captures the manager's type. They study quantity- and price-setting duopolies and show that hiring a relative profits-maximizing manager can serve as a commitment device and can, in fact, increase profit. For example, hiring managers who care about relative profits is advantageous under price competition with substitute goods. Under quantity competition, however, these managers behave more aggressively than pure profit-maximizing types and, consequently, profits decrease. Qualitatively, these results are in line with the insights obtained for the strategic incentives and the vertical separation settings. Englmaier (2011) studies an R&D tournament under price competition where firms can hire overoptimistic managers who make cost-reducing R&D investments. An overoptimistic manager type overestimates the quality of (and hence the consumers' willingness to pay for) the firm's product. In equilibrium, both firms hire overoptimistic managers and thereby avoid aggressive investments in R&D. Englmaier (2010) shows a similar result for (Cournot) quantity competition and demonstrates that hiring overoptimistic managers can even be profitable and welfare-enhancing. Yu (2014) generalizes this result for an n -firm oligopoly and shows that the relation between competition (measured by the number of firms in the market) and CEO overconfidence is inverted U-shaped. Englmaier and Reisinger (2014) consider a differentiated-products duopoly where firms can hire biased managers who under- or overestimate the size of the market. They show that in equilibrium both firms hire aggressive managers under price competition (managers underestimate demand) *and* under quantity competition (managers overestimate demand). In particular, the former result is in contrast with the literature on strategic incentives. Commitment might be also achieved by hiring managers with other-regarding or social preferences. For example, Kopel and Brand (2013) employ a strategic incentives model where a socially

concerned firm competes against a profit-maximizing rival. They show that the non-profit firm might strategically benefit from hiring an intrinsically motivated manager. In contrast to the profit-maximizing firm, the socially concerned firm might prefer to pay a straight salary to its manager instead of using strategic incentives based on profit and revenue. Finally, Cui, Raju, and Zhang (2007) introduce fairness concerns into a model of a manufacturer–retailer relation and demonstrate that a simple wholesale price can coordinate the supply chain.

4.2 Manager Types and Upstream Input Markets

So far we have looked at the influence of delegation or vertical separation on the strategic interaction between firms competing in the final product market. In what follows we (intentionally) abstract from final product market competition and try to see if commitment by delegating decision rights to a particular type of agent can provide other strategic benefits, e.g., if the firm deals with a supplier of an essential input.

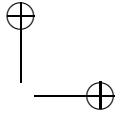
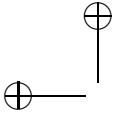
To be more specific, we consider a multi-product monopolist that is serving two independent markets, A and B. For the division that serves product market A the monopolist considers appointing a manager with a propensity to cause-related marketing, i.e., to donate a certain amount of the firm's sales to charity. It seems that unless this, e.g., increases the consumers' willingness to pay by reputation effects, hiring such a type of manager will decrease profits as it essentially "burns money". However, assume that an essential input is needed for manufacturing products A and B and that this input is purchased from an upstream monopoly supplier. Can it be a profit-enhancing strategy to hire such a type of socially concerned manager for division A? The answer is affirmative, and the intuition is as follows. If the manager of product division A donates to charity, marginal production costs are effectively increased. As a consequence, the production quantity offered in market A is decreased. This, in turn, decreases division A's demand for the essential input. In order to stimulate division A's demand for the input, the upstream supplier reduces its input price. Provided that the supplier cannot price discriminate between the input for products A and B, this input price reduction not only benefits division A but also division B. If the market size of product market B is sufficiently large, then total profits of the multi-product monopolist can indeed increase if it commits to hire a socially oriented type of manager for one product division. This result is rather unexpected if one compares it to the findings in the previous sections. Strategic delegation of the price or quantity choice to a charitable ("money-burning") type of manager cannot be beneficial if we just focus on downstream (Bertrand or Cournot) competition. It can, however, be a profitable commitment strategy against an upstream input supplier.

To elaborate on these ideas, assume that the multi-product monopolist offers quantities q_A and q_B in markets A and B. Inverse demands on the two markets are given by

$$p_A = a_A - q_A,$$

$$p_B = a_B - q_B.$$

The monopolist is buying the essential input for an input price of w from a supplier that produces with marginal costs $c = 0$ (for simplicity). Product division managers select the production quantities such that divisional profits are maximized. The monopolist considers hiring a socially concerned manager for product division A. This manager would donate an



amount z for each unit sold. In essence, this increases marginal production costs for product division A from w to $w + z$. Hence, the monopolist's total profit would be

$$\pi = \pi_A + \pi_B = [p_A q_A - (w + z) q_A] + [p_B q_B - w q_B].$$

The standard case of a non-philanthropic manager corresponds to $z = 0$. The question arises if it can be beneficial for the monopolist to hire a social manager with $z > 0$. The supplier profit is

$$\pi_S = w(q_A + q_B).$$

The timing of the game is as follows. First, the supplier selects the price w of the input. Second, division managers determine the quantities to maximize divisional profits. We will demonstrate that it is beneficial for the monopolist to select a socially concerned manager with $z > 0$ by comparing the resulting profit with the profit obtained with a non-philanthropic manager ($z = 0$). As an equilibrium concept we employ subgame-perfection.

Working backwards, product division i ($i = A, B$) first determines the (monopoly) quantity to maximize π_i . This yields

$$q_A = \frac{a_A - w - z}{2}$$

$$q_B = \frac{a_B - w}{2}.$$

The supplier anticipates these quantities and selects the input price w to maximize π_S . We obtain

$$w = \frac{a_A + a_B - z}{4}.$$

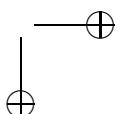
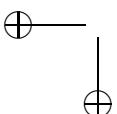
Note that the input price decreases with increasing z . Using the expression for the optimal supplier price, the resulting equilibrium total profit of the monopolist is

$$\begin{aligned} \pi = \pi_A + \pi_B &= (q_A)^2 + (q_B)^2 = \\ &= \left(\frac{3a_A - a_B - 3z}{8} \right)^2 + \left(\frac{-a_A + 3a_B + z}{8} \right)^2. \end{aligned}$$

We have $q_A \geq 0$ and $q_B \geq 0$ if and only if $a_A - 3a_B \leq z \leq \frac{3a_A - a_B}{3}$. Comparing the firm's profit with a socially concerned production division manager, $\pi(z > 0)$, with the firm's profit with a non-philanthropic division manager, $\pi(z = 0)$, yields

$$\pi(z > 0) > \pi(z = 0) \iff z \geq \frac{10a_A - 6a_B}{5}.$$

Consequently, if the donation per unit of quantity sold fulfills $\frac{10a_A - 6a_B}{5} \leq z \leq \frac{3a_A - a_B}{3}$, then (the solution is feasible and) the firm is better off hiring a socially concerned manager for its product division A. Observe that this happens if and only if $a_B \geq \frac{15}{13}a_A$, i.e., if the market B is sufficiently large compared to market A. Intuitively, committing to a manager who makes



donations based on sales raises marginal production costs. Obviously, this negatively affects the monopolist's total profits. However, the increase in marginal production costs in division A softens supplier pricing and, hence, decreases the input price. Although the profit of division A decreases, the increase in profit of division B can outweigh division A's profit reduction if market B is sufficiently large.

4.3 Organizational Design and the Firm's Supply Side

What this simple example shows is that decentralizing the upward channel (instead of the downward channel as in Section 3) can provide benefits. There are important strategic effects on the firm's supply side that affect firm behavior and profits even absent any strategic interaction on the final product market. Arya and Mittendorf (2010) give a comprehensive and insightful survey of the impact of supply-side considerations on organizational design when a firm is a seller or a buyer in input markets. Arya and Mittendorf (2015) study the supply-side consequences of market-based subsidies for socially desirable product market behavior. They find that subsidies for donated goods affect firm behavior upstream and downstream: supplier prices are decreasing, but profits are increasing; at the retail market firms have an incentive to raise output prices. Hence, such subsidies have undesirable spillover effects on retail prices. Arya and Mittendorf (2013) introduce a model of a multi-product firm that competes with a local (vertically integrated) rival firm in each of n market segments. The monopolist purchases the essential input that is used in each of the n products from a monopolist supplier. The authors show that it can be profitable for the monopolist to link its donations to charity to product sales as it softens supplier pricing. They derive the optimal number of segments $m^* \leq n$ where cause-related marketing should be introduced and study the impact of competition (measured by the level of products' substitutability) on m^* . Along the same lines as the agent appointment game we study in the previous subsection, their model provides a supply-side explanation for corporate social responsibility.

There are a number of contributions that study a firm's choice of organizational design in the presence of vertically related markets. In a recent paper, Kopel, Löffler, and Pfeiffer (2016) study the make-or-buy choice of a two-products monopolist when production of one product needs only a common input but the other product needs a common and a complementary specific input. They show that it might be optimal for the monopolist to deviate from an isolated least-cost comparison, i.e., the firm might decide to produce the specific input in-house even if marginal in-house production cost exceeds the per-unit input price of the supplier (and vice versa). This holds even under supplier price discrimination. Moreover, increased competition on the specific input market can result in a decrease of the firm's profit (due to a weakening of the commitment effect towards the common input supplier). Hinterecker and Kopel (2016) study the impact of asymmetric pollution tax rates if a multi-product monopolist can choose the location of production and sources an essential input from a global supplier. They show that if pollution taxes are increased in one country, a proportion of production is shifted to the low-tax country. The remaining production facility in the high-tax country works as a commitment mechanism to soften supplier pricing. As a consequence, a tax increase in one country can lead to an increase in the monopolist's profits, but also lead to an increase in total pollution in the high-tax country if pollution is transboundary. Kopel and Riegler (2016) compare the performance of two budgeting regimes, authoritative budgeting and participatory budgeting, and the impact of supply-side interactions. They

consider a multi-product firm that allocates cost budgets to each division. Under authoritative budgeting, the firm's headquarters determines the budget based on expected production costs. Under participative budgeting a manager submits a potentially biased cost report and, based on this report, headquarters determines the cost budget. In case of a biased cost report, the division manager consumes the slack. Kopel and Riegler show that slack induces a softening of supplier pricing and might increase the firm's profits. As a consequence, despite the fact that managerial slack burns money, participatory budgeting might be preferred to authoritative budgeting. Finally, Matsushima and Mizuno (2013) study a monopolistic firm that requires two complementary inputs. One input has to be purchased from a supplier. The other input can be made in-house, but the monopolist can also vertically separate and then purchase the other input from the separated unit. Wholesale prices are determined by Nash bargaining between the monopolist and the input supplier(s). Matsushima and Mizuno (2013) show that as the bargaining power of the independent supplier increases, the downstream monopolist tends to separate from its input unit. The rationale here is that separation creates an additional friction (due to double marginalization) that softens the pricing behavior of the independent supplier.

5 OTHER APPLICATIONS OF STRATEGIC DELEGATION IN OLIGOPOLY

In the literature there is a vast variety of applications that study strategic delegation in an oligopoly setting. In this section, we discuss just a narrow selection of papers that address key topics in industrial organization and game theory and have attracted sufficient attention of the scientific community up to date. At the end of the section, we also provide a few miscellaneous contributions on strategic delegation that we think are interesting and might lead to future developments.

One of the key topics in industrial economics is investments in R&D under quantity or price competition. The seminal paper of d'Aspremont and Jacquemin (1988) on the incentives for cost-reducing investments if R&D spillovers exist has been extended by Zhang and Zhang (1997). They assume that firms delegate the task of R&D investments and quantity choice to their managers who are given strategic incentives based on profits and revenues. Kopel and Riegler (2006) argue that closed-form solutions of this R&D game under positive R&D spillovers cannot be obtained and provide some numerical examples. More recent papers study investments in process innovation under strategic delegation but drop the assumption of R&D spillovers. Veldman et al. (2014) study a duopoly R&D investment game where managerial compensation is based on profit and a bonus for cost reductions. They demonstrate that the main insights of the strategic incentives literature hold but also show that if the R&D efficiency of the firms is asymmetric, then the firm with a sufficiently high R&D efficiency can actually obtain higher profits than in the no delegation case. See also Overvest and Veldman (2008) for a similar analysis. Mitrokostas and Petrakis (2014) study a strategic delegation game where the manager can either determine both R&D investments and quantity (full delegation) or R&D investments are determined by the firm and the manager can only determine the quantity (partial delegation). They provide conditions under which full delegation is optimal (e.g., if initial marginal costs are high) or asymmetric delegation, i.e., partial delegation for one firm and full delegation for the other, emerges in equilibrium. They also demonstrate that under

price competition, universal partial delegation may occur. Finally, Mahathi, Pal, and Ramani (2016) study technology adoption if price or quantity decisions are contractually delegated to a manager, but owners take decisions regarding the adoption date of a new technology. They demonstrate that the date of technology adoption under delegation differs from the no-delegation case. Moreover, they find that early or late adoption is influenced by the mode of competition (Bertrand or Cournot) and by the performance measure (absolute or relative) used in managers' compensation contracts.

Another hotly debated topic in the literature is (horizontal) mergers between firms and the resulting effects on firms' outputs and profits and welfare. One of the puzzling early results in the literature (which contrasts empirical evidence) is that a merger is privately profitable only if a relatively high fraction of firms in a Cournot market engage in the merger. González-Maestre and López-Cuñat (2001) consider strategic delegation in a Cournot oligopoly with linear demand and cost functions and show that with delegation the required number of firms for a merger to be profitable is substantially reduced compared to the non-delegation case. Ziss (2001) confirms this insight in a more general Cournot oligopoly model. Straume (2006) studies strategic delegation in a three-firm Cournot industry and finds that there is always a conflict between private and social merger incentives under delegation. Further work on strategic delegation and mergers is done by Ziss (2007) and Creane and Davidson (2004).

There are a number of miscellaneous topics that are studied through the lens of strategic delegation. We would like to discuss a few papers we find interesting. Some of these contributions investigate settings where *ex ante* symmetric firms choose asymmetric strategies in equilibrium. We begin with Mujumdar and Pal (2007), who study optimal strategic incentives in a dynamic, two-period quantity-setting homogeneous duopoly. Both firms make their production decisions in each of the two periods, while market-clearing occurs only after the end of period two. They find that firm heterogeneity might arise as an endogenous equilibrium outcome. Both owners delegate the production decision but the early-moving firm writes a profit-based incentive contract for the manager, whereas the late-moving firm bases the incentive contract only on sales revenues. Additionally, they show that the leading firm may have a first-mover advantage. Kopel and Löffler (2012) also study the endogenous emergence of leader–follower roles and asymmetric organizational structures of *a priori* symmetric firms and find asymmetric equilibria in a price- and quantity-setting environment. Kopel and Löffler (2008) study equilibrium combinations of multiple commitment strategies – investment in cost-reducing R&D, strategic delegation of decision rights to managers, and the role of Stackelberg leader or follower at the quantity stage. They find that there is a unique equilibrium in which both firms invest in process R&D, only the follower delegates, and the follower can overcome the first-mover advantage of the quantity leader and obtain a higher profit than the leader. Vroom (2006) investigates the use of strategic incentives and organizational design and finds that joint employment can reduce competitive rivalry. He further finds that in equilibrium one firm might choose decentralization and uses strategic incentives to make its manager more aggressive while the other firm selects centralization and softens managerial behavior.

Bagnoli and Watts (2015) combine strategic incentives in a Cournot duopoly with an *ex ante* disclosure setting. Since contractually delegating the output choice to a manager has a commitment effect only if this is observable by the rival firm, it is a natural question to ask how (strategic) disclosure and incentives interact. In their model, owners jointly delegate

production and disclosure choices to managers and write incentive contracts based on profit and sales revenue. Managers commit to ex ante disclose any private information to its rival. They find that the anticipation of disclosure by either manager substitutes for contractual incentives to compete more aggressively in the product market. They also show that owners who provide contractual incentives to their managers to act more aggressively at the same time give their managers incentives to disclose information that is more useful in learning about the rival's costs than the own firm's costs. The rival firm instead provides incentives to the manager to disclose information that is more useful in understanding the own firm's costs. In any event, their results demonstrate that there is an important interaction between contractual incentives chosen by firms and the competing firms' information environments. Etro (2011) criticizes the lack of robustness of results on the strategic use of contracts with regard to the mode of competition. The usual result under price competition (when the number of firms is exogenous) is that the managers are punished for sales to induce them to keep prices high. However, the resulting high profits would attract entry. Etro shows, for example, that when the number of firms in the market is endogenous, then strategic delegation with positive sales incentives for managers is optimal under both quantity and price competition. Hoernig (2012) introduces network effects in a linear Bertrand duopoly and demonstrates a similar result. He shows that if network effects are sufficiently strong, then in equilibrium owners want their managers to be aggressive. The reason here is that in this case incentive rates become strategic substitutes and owners have strong incentives to fight for market share. Chirco and Scrimitore (2013) study endogenous price or quantity choice in a differentiated duopoly with network effects. They find (in contrast to Singh and Vives, 1984) that under strategic delegation and sufficiently strong network effects, there is a unique equilibrium where both firms select a price contract. Finally, Kräkel (2005) considers tournament competition with and without noise and studies the strategic use of linear contracts based on profits and sales. He shows that the outcomes sharply differ from the findings for linear oligopoly in that asymmetric equilibria exist where one owner puts a positive weight on sales, whereas the other owner chooses a negative weight.

6 CONCLUDING REMARKS

In this chapter we have discussed various streams of the literature that have shown that delegation of decision rights to agents with divergent objectives can yield strategic benefits for the delegating party in oligopolistic markets. The strategic incentives approach presented in Section 2 shows that a firm's profit can be increased by delegating, e.g., quantity, price, or R&D investment decisions, to a manager who is given appropriate incentives by a compensation contract based on output, sales revenue, or market share in addition to profit. The vertical separation approach presented in Section 3 demonstrates that a manufacturer's profit can be increased by vertically separating from an independent retail unit that determines production quantity or market price to maximize retail profit. Likewise, upstream decentralization or strategic outsourcing to an independent supplier of an input can provide benefits under oligopolistic competition. Finally, the agent appointment approach presented in Section 4 demonstrates that delegating decisions to a particular type of manager – biased, overoptimistic, or socially concerned – can yield strategic benefits for the delegating party. This can occur if product market competition is oligopolistic, but can also occur on

the supply side if a firm's commitment or internal frictions within the firm impact supplier (pricing) behavior.

We end the chapter with two suggestion for further research. First, we believe that an interesting topic is the strategic impact of psychological and sociological aspects of human decision-making behavior under oligopolistic competition. In behavioral game theory and behavioral agency theory issues like, e.g., fairness, reciprocity, envy and status-concerns, identity and the role of norms, are intensively discussed (see Kopel and Brand, 2013 for references). Experimental evidence demonstrates that systematic deviations of subjects from the predictions of game and agency models might occur. The costs of intra-firm and inter-firm comparisons between individuals in organizations might even impact firms' boundaries (Nickerson and Zenger, 2008). We suggest that a more systematic and thoughtful study of the influence of these aspects in a strategic delegation setting might show the robustness of existing findings. Additionally, it might also provide an explanation for real-world practices and might lead to new (potentially testable) results.

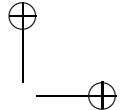
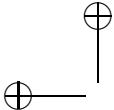
Second, we further believe that the combination of strategic delegation with environmental issues under oligopolistic interaction is another interesting topic to pursue (see Lambertini, 2013). Regulators select tax policies, firms choose locations and design their organizational structures, and managers are compensated based on their firms' environmental performances. Additionally, suppliers make investments in green technology and jointly with the firms' responsible behaviors and investments affect consumers' willingness to pay. All these choices interact and set the stage for oligopolistic competition in the global market with multiple products and suppliers. Again, we suggest that a more systematic study of these aspects might lead to interesting insights into the practices of green innovation, the recent increase of investments in corporate social responsibility, and the associated discussion about proper performance measures in managerial compensation contracts to achieve green goals. All parties – regulators, firms, suppliers, and consumers – commit to their individually optimal strategy. However, firms' profits, suppliers' profits, consumer welfare, and total welfare are determined by the interaction of these commitment strategies in oligopolistic markets.

REFERENCES

- Aggarwal, R.K. and A.A. Samwick (1999), Executive Compensation, Strategic Competition, and Relative Performance Evaluation: Theory and Evidence, *Journal of Finance* 54(6), 1999–2043.
- Alles, M. and S. Datar (1998), Strategic Transfer Pricing, *Management Science* 44(4), 451–461.
- Anderson, E.J. and Y. Bao (2010), Price Competition With Integrated and Decentralized Supply Chains, *European Journal of Operational Research* 200(1), 227–234.
- Arya, A. and B. Mittendorf (2010), Input Markets and the Strategic Organization of the Firm, *Foundations and Trends in Accounting* 5(1), 1–97.
- Arya, A. and B. Mittendorf (2013), A Supply-side Explanation for the Use of Cause Marketing, *Working Paper*.
- Arya, A. and B. Mittendorf (2015), Supply Chain Consequences of Subsidies for Corporate Social Responsibility, *Production and Operations Management* 24(8), 1346–1357.
- Asseburg, H. and C. Hofmann (2010), Relative Performance Evaluation and Contract Externalities, *OR Spectrum* 32(1), 1–20.
- Baggs, J. and J.-E. de Bettignies (2007), Product Market Competition and Agency Costs, *Journal of Industrial Economics* 55(2), 289–323.
- Bagnoli, M. and S.G. Watts (2015), Delegating Disclosure and Production Choices, *The Accounting Review* 90(3), 835–857.
- Bagwell, K. (1995), Commitment and Observability in Games, *Games and Economic Behavior* 8, 271–280.

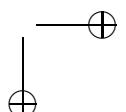
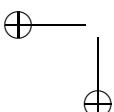
- Balasubramanian, S. and P. Bhardwaj (1994), When Not All Conflict is Bad: Manufacturing–Marketing Conflict and Strategic Incentive Design, *Management Science* 50(4), 489–502.
- Barreda-Tarazona, I., N. Georgantzis, and C. Manasaki et al. (2016), Endogenous Managerial Compensation Contracts in Experimental Quantity-setting Duopolies, *Economic Modelling* 54, 205–217.
- Bassi, M., M. Pagnozzi, and S. Piccolo (2015), Product Differentiation by Competing Vertical Hierarchies, *Journal of Economics and Management Strategy* 24(4), 904–933.
- Basu, K. (1995), Stackelberg Equilibrium in Oligopoly: An Explanation Based on Managerial Incentives, *Economics Letters*, 49, 459–464.
- Baye, M., K.J. Crocker, and J. Ju (1996), Divisionalization, Franchising, and Divestiture Incentives in Oligopoly, *American Economic Review* 86(1), 223–236.
- Berr, F. (2011), Stackelberg Equilibria in Managerial Delegation Games, *European Journal of Operational Research* 212, 251–262.
- Bhardwaj, P. (2001), Delegating Pricing Decisions, *Marketing Science* 20(2), 143–169.
- Bhardwaj, P. and S. Balasubramanian (2005), Managing Channel Profits: The Role of Managerial Incentives, *Quantitative Marketing and Economics* 3, 247–279.
- Bonanno, G. and J. Vickers (1988), Vertical Separation, *The Journal of Industrial Economics* 36(3), 257–265.
- Bowley, A.L. (1924), *The Mathematical Groundwork of Economics*, Oxford: Oxford University Press.
- Bulow, J.I., J.D. Geanakoplos, and P.I. Klemperer (1985), Multimarket Oligopoly: Strategic Substitutes and Complements, *Journal of Political Economy* 93(3), 488–511.
- Caillaud, B. and P. Rey (1995), Strategic Aspects of Vertical Delegation, *European Economic Review* 39, 421–431.
- Casadesus-Masanell, R. (2004), Trust in Agency, *Journal of Economics and Management Strategy* 13(3), 375–404.
- Chalioti, E. (2015), Incentive Contracts Under Product Market Competition and R&D Spillovers, *Economic Theory* 58(2), 305–328.
- Chirco, A. and M. Scrimitore (2013), Choosing Price or Quantity? The Role of Delegation and Network Externalities, *Economics Letters* 121, 482–486.
- Choi, S.C. (1991), Price Competition in a Channel Structure With a Common Retailer, *Marketing Science* 10, 271–296.
- Choi, S.C. (1996), Price Competition in a Duopoly Common Retailer Channel, *Journal of Retailing* 72, 117–134.
- Chou, C.-H. (2014), Strategic Delegation and Vertical Integration, *Managerial and Decision Economics* 35(8), 580–586.
- Corts, K.S. and D.V. Neher (2003), Credible Delegation, *European Economic Review* 47, 395–407.
- Coughlan, A.T. and B. Wernerfelt (1989), On Credible Delegation by Oligopolists: A Discussion of Distribution Channel Management, *Management Science* 35(2), 226–239.
- Creane, A. and C. Davidson (2004), Multidivisional Firms, Internal Competition, and the Merger Paradox, *Canadian Journal of Economics* 37(4), 951–977.
- Cui, T.H., J.S. Raju and Z.J. Zhang (2007), Fairness and Channel Coordination, *Management Science* 53(8), 1303–1314.
- Cuñat, V. and M. Guadalupe (2005), How Does Product Market Competition Shape Incentive Contracts? *Journal of the European Economic Association* 3(5), 1058–1082.
- D'Aspremont and Jacquemin (1988), Cooperative and Noncooperative R&D in a Duopoly With Spillovers, *American Economic Review* 78, 1133–1137.
- De Bettignies, J.-E. (2006), Product Market Competition and the Boundaries of the Firm, *Canadian Journal of Economics* 39(3), 948–970.
- Dockner, E.J. and C. Löffler (2015), Rivalry Restraint as Equilibrium Behavior, *Journal of Economics and Management Strategy* 24(1), 189–209.
- Du, N., J.S. Heywood, and G. Ye (2013), Strategic Delegation in an Experimental Mixed Duopoly, *Journal of Economic Behavior and Organization* 87, 91–100.
- Dürr, O.M. and R.F. Göx (2011), Strategic Incentives for Keeping One Set of Books in International Transfer Pricing, *Journal of Economics and Management Strategy* 20(1), 269–298.
- Englmaier, F. (2010), Managerial Optimism and Investment Choice, *Managerial and Decision Economics* 31(4), 303–310.
- Englmaier, F. (2011), Commitment in R&D Tournaments via Strategic Delegation to Overoptimistic Managers, *Managerial and Decision Economics* 32(1), 63–69.
- Englmaier, F. and M. Reisinger (2014), Biased Managers as Strategic Commitment, *Managerial and Decision Economics* 35(5), 350–356.
- Etro, F. (2011), Endogenous Market Structures and Contract Theory: Delegation, Principal–Agent Contracts, Screening, Franchising and Tying, *European Economic Review* 55(4), 463–479.
- Fershtman, C. (1985), Managerial Incentives as a Strategic Variable in Duopolistic Environment, *International Journal of Industrial Organization* 3, 245–253.
- Fershtman, C. and U. Gneezy (2001), Strategic Delegation: An Experiment, *RAND Journal of Economics* 32(2), 352–368.

- Fershtman, C. and K.L. Judd (1987), Equilibrium Incentives in Oligopoly, *American Economic Review* 77, 927–940.
- Fershtman, C. and K.L. Judd (1990), Strategic Incentive Manipulation in Rivalrous Agency, *Working Paper*.
- Fershtman, C. and E. Kalai (1997), Unobservable Delegation, *International Economic Review* 38(4), 763–774.
- Fershtman, C., K.L. Judd, and E. Kalai (1991), Observable Contracts: Strategic Delegation and Cooperation, *International Economic Review* 32(3), 551–559.
- Fudenberg, D. and J. Tirole (1984), The Fat-cat Effect, The Puppy-dog Ploy, and the Lean and Hungry Look, *American Economic Review* 74(2), 361–366.
- Gal-Or, E. (1997), Multiprincipal Agency Relationships as Implied by Product Market Competition, *Journal of Economics and Management Strategy* 6(2), 235–256.
- González-Maestre, M. (2000), Divisionalization and Delegation in Oligopoly, *Journal of Economics and Management Strategy* 9(3), 321–338.
- González-Maestre, M. and J. López-Cuñat (2001), Delegation and Mergers in Oligopoly, *International Journal of Industrial Organization* 19, 1263–1279.
- Göck, R.F. (2000), Strategic Transfer Pricing, Absorption Costing, and Observability, *Management Accounting Research* 11, 327–348.
- Göck, R.F. and U. Schiller (2007), An Economic Perspective on Transfer Pricing, in C.S. Chapman, A.G. Hopwood, and M.D. Shields (eds), *Handbook of Management Accounting Research*, Amsterdam: Elsevier, pp. 673–695.
- Graziano, C. and B.M. Parigi (1998), Do Managers Work Harder in Competitive Industries? *Journal of Economic Behavior and Organization* 34, 489–498.
- Gupta, S. and R. Loulou (1998), Process Innovation, Product Differentiation and Channel Structure: Strategic Incentives in a Duopoly, *Management Science* 17(4), 301–316.
- Heifetz, A., C. Shannon, and Y. Spiegel (2007), What To Maximize If You Must, *Journal of Economic Theory* 133(1), 31–57.
- Hinterecker, H. and M. Kopel (2016), Supply Side Effects of Pollution Tax Asymmetries, *Working Paper*.
- Hoernig, S. (2012), Strategic Delegation Under Price Competition and Network Effects, *Economics Letters* 117, 487–489.
- Huang, R., C. Marquardt, and B. Zhang (2015), Using Sales as a Performance Measure, *Working Paper*, available at <http://ssrn.com/abstract=2636950>.
- Huck, S., W. Müller, and H.-T. Normann (2004), Strategic Delegation in Experimental Markets, *International Journal of Industrial Organization* 22, 561–574.
- Irmen, A. (1998), Precommitment in Competing Vertical Chains, *Journal of Economic Surveys* 12(4), 333–359.
- Irwin, D.A. (1991), Mercantilism as Strategic Trade Policy: The Anglo-Dutch Rivalry for the East India Trade, *Journal of Political Economy* 99(6), 1296–1314.
- Jansen, T., A. van Lier, and A. van Witteloostuijn (2007), A Note on Strategic Delegation: The Market Share Case, *International Journal of Industrial Organization* 25, 531–539.
- Jansen, T., A. van Lier and A. van Witteloostuijn (2009), On the Impact of Managerial Bonus Systems on Firm Profit and Market Competition: The Cases of Pure Profit, Sales, Market Share and Relative Profits Compared, *Managerial and Decision Economics* 30, 141–153.
- Karuna, C. (2007), Industry Product Market Competition and Managerial Incentives, *Journal of Accounting and Economics* 43, 275–297.
- Katz, M.L. (1991), Game-playing Agents: Unobservable Contracts as Precommitments, *RAND Journal of Economics* 22(3), 307–328.
- Katz, M.L. (2006), Observable Contracts as Commitments: Interdependent Contracts and Moral Hazard, *Journal of Economics and Management Strategy* 15(3), 685–706.
- Kedia, S. (2006), Estimating Product Market Competition: Methodology and Application, *Journal of Banking and Finance* 30, 875–894.
- Koçkesen, L. (2007), Unobservable Contracts as Precommitments, *Economic Theory* 31, 539–552.
- Koçkesen, L. and E.A. Ok (2004), Strategic Delegation by Unobservable Incentive Contracts, *Review of Economic Studies* 71, 397–424.
- Koçkesen, L., E.A. Ok, and R. Sethi (2000), The Strategic Advantage of Negatively Interdependent Preferences, *Journal of Economic Theory* 92, 274–299.
- Königstein, M. and Müller, W. (2001), Why Firms Should Care for Customers, *Economics Letters* 72(1), 47–52.
- Kopel, M. and B. Brand (2013), Why Do Socially Concerned Firms Provide Low-powered Incentives to Their Managers? *Working Paper*.
- Kopel, M. and F. Lamantia (2016), Mixed Industry Outcomes in Oligopoly Markets with Socially Concerned Firms, *Working Paper*.
- Kopel, M. and L. Lambertini (2013), On Price Competition With Market Share Delegation Contracts, *Managerial and Decision Economics* 34(1), 40–43.
- Kopel, M. and C. Löffler (2008), Commitment, First-mover, and Second-mover Advantage, *Journal of Economics* 94(2), 143–166.

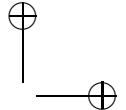
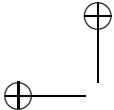


284 *Handbook of game theory and industrial organization: applications*

- Kopel, M. and C. Löffler (2012), Organizational Governance, Leadership, and the Influence of Competition, *Journal of Institutional and Theoretical Economics* 168, 362–392.
- Kopel, M. and C. Riegler (2006), R&D in a Strategic Delegation Game Revisited: A Note, *Managerial and Decision Economics* 27(7), 605–612.
- Kopel, M. and C. Riegler (2016), Slack and Participative Budgeting – New Aspects of the Benefits of Slack Building, *Working Paper*.
- Kopel, M., F. Lamantia, and F. Szidarovszky (2014), Evolutionary Competition in a Mixed Market With Socially Concerned Firms, *Journal of Economic Dynamics & Control* 48, 394–409.
- Kopel, M., C. Löffler, and T. Pfeiffer (2016), Sourcing Strategies of a Multi-input-Multi-product Firm, *Journal of Economic Behavior and Organization* 127, 30–45.
- Kopel, M., M. Pezzino, and A. Ressi (2016), Contract Bargaining and Location Choice, *Managerial and Decision Economics* 37(2), 140–148.
- Kräkel, M. (2005), Strategic Delegation in Oligopolistic Tournaments, *Review of Economic Design* 9, 377–396.
- Krishna, P. (2001), On Competition and Endogenous Firm Efficiency, *Economic Theory* 18, 753–760.
- Lambertini, L. (2013), Oligopoly, the Environment and Natural Resources, London and New York: Routledge.
- Larcker, D. and B. Tayan (2011), *Corporate Governance Matters*, Upper Saddle River, NJ: Pearson.
- Liu, Y. and R.K. Tyagi (2011), The Benefits of Competitive Upward Channel Decentralization, *Management Science* 57(4), 741–751.
- Maggi, G. (1999), The Value of Commitment with Imperfect Observability and Private Information, *RAND Journal of Economics* 30(4), 555–574.
- Mahathi, A., R. Pal, and V. Ramani (2016), Competition, Strategic Delegation, and Delay in Technology Adoption, *Economics of Innovation and New Technology* 25(2), 143–171.
- Martinez-Giralt, X. and D.J. Neven (1988), Can Price Competition Dominate Market Segmentation, *Journal of Industrial Economics* 36(4), 431–442.
- Matsui, K. (2012), Strategic Upfront Marketing Channel Integration as an Entry Barrier, *European Journal of Operational Research* 220(3), 865–875.
- Matsushima, N. and T. Mizuno (2013), Vertical Separation as a Defense Against Strong Suppliers, *European Journal of Operational Research* 228, 208–216.
- Merzoni, G. (2000), Strategic Delegation in Cournot Oligopoly With Incomplete Information, in M.R. Baye (ed.), *Industrial Organization*, Vol. 9, Bingley, UK: Emerald, pp. 279–305.
- Miller, N. and A. Pazgal (2001), The Equivalence of Price and Quantity Competition with Delegation, *RAND Journal of Economics* 32, 284–301.
- Miller, N. and A. Pazgal (2002), Relative Performance as a Strategic Commitment Mechanism, *Managerial and Decision Economics* 23, 51–68.
- Milliou, C. and E. Petrakis (2007), Upstream Horizontal Mergers, Vertical Contracts, and Bargaining, *International Journal of Industrial Organization* 25, 963–987.
- Mitrokostas, E. and E. Petrakis (2014), Organizational Structure, Strategic Delegation and Innovation in Oligopolistic Industries, *Economics of Innovation and Technology* 23(1), 1–24.
- Moorthy, K.S. (1988), Strategic Decentralization in Channels, *Marketing Science* 7(4), 335–355.
- Mujumdar, S. and D. Pal (2007), Strategic Managerial Incentives in a Two-period Cournot Duopoly, *Games and Economic Behavior* 58, 338–353.
- Nickerson, J.A. and T.R. Zenger (2008), Envy, Comparison Costs, and the Economic Theory of the Firm, *Strategic Management Journal* 29(13), 1429–1449.
- O'Brien, D.P. and G. Shaffer (1992), Vertical Control With Bilateral Contracts, *RAND Journal of Economics* 23(3), 299–308.
- Overvest, B.M. and J. Veldman (2008), Managerial Incentives for Process Innovation, *Managerial and Decision Economics* 29, 539–545.
- Pagnozzi, M. and S. Piccolo (2012), Vertical Separation With Private Contracts, *Economic Journal* 122, 173–207.
- Piccolo, S., M. D'Amato, and R. Martina (2008), Product Market Competition and Organizational Slack Under Profit-target Contracts, *International Journal of Industrial Organization* 26, 1389–1406.
- Plehn-Dujowich, J.M. and K. Serfes (2010), Strategic Managerial Compensation Arising From Product Market Competition, available at <http://ssrn.com/abstract=1540629>.
- Raiith, M. (2003), Competition, Risk, and Managerial Incentives, *American Economic Review* 93(4), 1425–1436.
- Reitman, D. (1993), Stock Options and the Strategic Use of Managerial Incentives, *American Economic Review* 83(3), 513–524.
- Rey, P. and J. Stiglitz (1988), Vertical Restraints and Producers' Competition, *European Economic Review* 32, 561–568.
- Rey, P. and J. Stiglitz (1995), The Role of Exclusive Territories in Producers' Competition, *RAND Journal of Economics* 26(3), 431–451.
- Rey, P. and T. Vergé (2008), Economics of Vertical Restraints, in P. Buccrossi (ed.), *Handbook of Antitrust Economics*, Cambridge, MA: MIT Press, pp. 353–390.



- Ritz, R.A. (2008), Strategic Incentives for Market Share, *International Journal of Industrial Organization* 26, 586–597.
- Rotemberg, J.J. (1994), Human Relations in the Workplace, *Journal of Political Economy* 102(4), 684–717.
- Rysman, M. (2001), How Many Franchises in a Market? *International Journal of Industrial Organization* 19, 519–542.
- Salas Fumás, V.S. (1992), Relative Performance Evaluation of Management. The Effects on Industrial Competition and Risk Sharing, *International Journal of Industrial Organization* 10, 473–489.
- Schelling, T.C. (2006), *Strategies of Commitment*, Cambridge, MA: Harvard University Press.
- Schmidt, K.M. (1997), Managerial Incentives and Product Market Competition, *Review of Economic Studies* 64(2), 191–213.
- Sengul, M. and J. Gimeno (2013), Constrained Delegation: Limiting Subsidiaries' Decision Rights and Resources in Firms That Compete Across Multiple Industries, *Administrative Science Quarterly* 58(3), 420–471.
- Sengul, M., J. Javier Gimeno and J. Dial (2012), Strategic Delegation: A Review, Theoretical Integration, and Research Agenda, *Journal of Management* 38, 375–414.
- Singh, N. and X. Vives (1984), Price and Quantity Competition in a Differentiated Duopoly, *RAND Journal of Economics* 15, 546–554.
- Sklivas, S.D. (1987), The Strategic Choice of Managerial Incentives, *RAND Journal of Economics* 18, 452–458.
- Slade, M. (1998), Strategic Motives for Vertical Separation: Evidence From Retail Gasoline Markets, *Journal of Law, Economics, and Organization* 14(1), 84–113.
- Spagnolo, G. (2000), Stock-related Compensation and Product-market Competition, *RAND Journal of Economics* 31(1), 22–42.
- Straume, O.R. (2006), Managerial Delegation and Merger Incentives With Asymmetric Costs, *Journal of Institutional and Theoretical Economics* 162, 450–469.
- Tabuchi, T. (2012), Multiproduct Firms in Hotelling's Spatial Competition, *Journal of Economics & Management Strategy* 21(2), 445–467.
- Van Witteloostuijn A., T. Jansen, and A. van Lier (2007), Bargaining Over Managerial Contracts in Delegation Games: Managerial Power, Contract Disclosure and Cartel Behavior, *Managerial and Decision Economics* 28(8), 897–904.
- Veldman, J., W. Klingenberg, G.J.C. Gaalman, and R.H. Teunter (2014), Getting What You Pay For – Strategic Process Improvement Compensation and Profitability Impact, *Production and Operations Management* 23(8), 1387–1400.
- Vickers, J. (1985), Delegation and the Theory of the Firm, *Economic Journal*, Supplement 95, 138–147.
- Vroom, G. (2006), Organizational Design and the Intensity of Rivalry, *Management Science* 52(11), 1689–1702.
- Vroom, G. and J. Gimeno (2007), Ownership Form, Managerial Incentives, and the Intensity of Rivalry, *Academy of Management Journal* 50(4), 901–922.
- Yu, C.-F. (2014), CEO Overconfidence and Overinvestment Under Product Market Competition, *Managerial and Decision Economics* 35(8), 574–579.
- Zhang, J. and Z. Zhang (1997), R&D in a Strategic Delegation Game, *Managerial and Decision Economics* 18, 391–398.
- Ziss, S. (1998), Divisionalization and Product Differentiation, *Economics Letters* 59, 133–138.
- Ziss, S. (2001), Horizontal Mergers and Delegation, *International Journal of Industrial Organization* 19, 471–492.
- Ziss, S. (2007), Hierarchies, Intra-firm Competition and Mergers, *International Journal of Industrial Organization* 25, 237–260.



11. Platforms and network effects

*Paul Belleflamme and Martin Peitz**

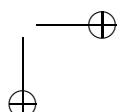
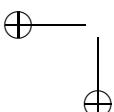
1 INTRODUCTION

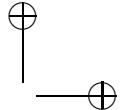
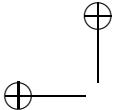
Over the last two decades, the fast penetration of the Internet and the digitization of information products has led to the rise of electronic intermediaries such as Amazon, Google, and Facebook. Some of these intermediaries have become darlings on the stock market, reflecting the belief that they have or will become central players in market economies. Of course, intermediaries are not a phenomenon of the Internet, but have been around since ancient times. Many of these intermediaries play an important role because of supply-side or demand-side scale effects. Of particular importance are the latter, as many intermediaries require a certain volume of usage to attract additional users. This puts network effects at the core of intermediaries or, as we will call them, platforms.

A platform brings together a typically large number of users who interact with each other. For instance, the traditional telecommunications provider is such a platform that brings together people who may want to engage in communicating with each other. The more users on the network the more valuable the communication service. This is an example of direct network effects.

On many platforms we can distinguish between distinct groups of users, whose activities affect the well-being of those in another group. One example are software platforms: they bring together application developers and end users. Here, everything else given, end users may not care about the presence of other users, but only about the number and quality of application developers, while developers only care about the number and demand of end users. In this case, network effects are indirect, as end users care about participation and usage of other end users only indirectly, as more end users attract more developers, which is beneficial for each end user. The platform managing the interaction among distinct groups of consumers is called two-sided. Some platforms allow for the interaction of buyers and sellers. Shopping malls are an example, as they offer retail space to sellers and invite buyers to go shopping. Everything else given, sellers prefer a shopping mall that attracts more buyers and buyers prefer a shopping mall that hosts more sellers. Trade fairs, flea markets, auction houses, and yellow pages have similar features. While some of these platforms have been around for centuries, platforms as a way to organize market activities have arguably gained more prominence with the rise of the Internet. To enable consumers to choose among a myriad of offerings, horizontal and vertical search engines as well as price search engines, booking portals, online auction and retail places have become commonplace. As these digital platforms are not subject to physical capacity constraints and can quickly guide a potential buyer to products of interest, they are able to manage huge volumes of interactions between buyers and sellers.

* We thank Markus Reisinger for helpful comments. Martin Peitz gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft (PE 813/2-2).





Platforms try to manage user participation and volumes of interaction. They can use price and non-price instruments for this purpose. In particular, they may court one particular group of users, e.g., buyers, to extract revenues from another group of users, e.g., sellers, who see the efforts of the platform to attract the first group of users. Managing users' participation and volume of interaction thus depends on the ability of platforms to convince users about the decisions taken by other users. Users can become convinced because of past decisions other users have made (to the extent that they are not easily revised) or by a platform's actions such as publicly observed prices applicable to other users, which in turn affect expectations about the decisions these other users will make.

This chapter reviews some key contributions to the economics of network effects and two-sided markets. In Section 2 we provide a discussion of network effects, criteria to classify different platform markets, and a number of examples. We explore the economics of markets with network effects in Section 3, and of two-sided markets in Section 4. Our aim is not to provide an exhaustive overview of the literature on network effects and two-sided markets; instead, we pick a few contributions and elaborate on some key findings in a number of stylized settings.¹

2 NETWORK EFFECTS AND PLATFORMS: A FIRST LOOK

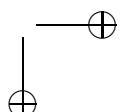
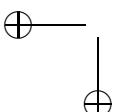
In this section we give a definition of network effects, drawing a distinction between direct (within-group) and indirect (cross-group) effects. We also explain what we mean by 'platforms'. We then offer some illustrations.

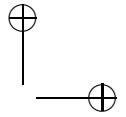
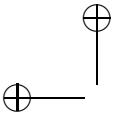
2.1 Defining Network Effects

Network effects are present if users care about participation and usage decisions of other users. The group of users making the same usage decision is loosely called the 'network'. In the simplest setting, users care from an *ex ante* perspective only about the *size* of the network; in a more general setting, also the *identity* of the users of the network matters. For instance, social norms, languages and communication devices clearly generate positive network effects: the more they are adopted, the larger the utility they confer to their adopters. For a social norm, only the size of the network matters; in contrast, for communication means (e.g., a particular language or an instant messaging application), a user is primarily concerned with the decision of the subset of users with whom she has regular interactions; hence, the identity of the users of the network matters.

Network effects may emerge in a large variety of contexts and may be positive or negative depending on the circumstances. Road congestion and traffic jams are the prototypical examples of negative network effects: the more drivers choose a particular road at a particular moment, the slower the traffic on that road at that moment and, thereby, the lower the utility of every driver. Fashion and fads generate positive network effects for those individuals whose utility increases when they conform with the choices of others. Yet, the exact opposite applies for snobs, who value the idea of having different tastes than the 'mass': for them, having someone choosing like them generates a negative network effect – see, e.g., Grilo, Shy, and

¹ To this end, we draw from and expand upon Chapters 20, 21, and 22 of Belleflamme and Peitz (2015).





Thisse (2001). Another example of positive network effects can be found in the choice of geographical locations by firms. Following the seminal work of Marshall (1890 [1920]), the economic geography literature explains why firms can benefit from locating close to one another; one explanation is that when more firms locate in the same region, more workers (or, more generally, input suppliers) are drawn to this region, which in turn makes the region more attractive for firms.² We see in this last example that network effects do not arise directly from the firms making the same choice but indirectly through the induced decisions of another group of agents (i.e., workers and/or input suppliers).

To clarify what we will discuss in this chapter, we first distinguish between the direct and indirect sources of network effects; we then narrow our focus on so-called ‘platforms’, which somehow manage network effects.

2.1.1 Within-group and cross-group external effects

The number of participants and the intensity of use may affect a population of agents, as argued above. In many environments, we can distinguish different groups. One example is trading platforms on which buyers and sellers interact. In this case, positive cross-group external effects are present because buyers are, everything else given, better off the more sellers are present and vice versa. Another example is content platforms that carry advertising. Here, vertically integrated content is offered to entice consumers to join the platform. Consumers then may pay directly for participation or indirectly with their attention to advertising, which is bundled with content. If consumers’ utility is decreasing in the volume of advertising, advertisers exert a negative cross-group external effect on consumers, while consumers exert a positive cross-group external effect on advertisers. In addition to cross-group effects, also within-group external effects may be present. For instance, if the platform becomes congested if too many actors from one group participate, there are negative within-group external effects. Also, increased competition between sellers constitutes a negative within-group external effect on the seller side.

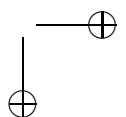
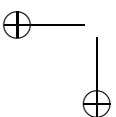
2.1.2 Markets with platforms

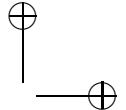
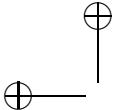
Markets with platforms can be broadly defined as *markets where the interaction between at least some participants is facilitated and managed by an intermediary*. Managing this interaction can take many forms; the most obvious ways are setting prices for participation or usage, or setting participation levels. The intermediary may have other instruments at its disposal. For instance, it may impose certain contractual terms, and it may provide monetary or non-monetary benefits for certain actions.

Our definition includes markets with a single group that exhibits within-group external effects, as long as an intermediary has at least one instrument to affect these within-group external effects and market outcomes. This sets our definition apart from recent definitions entering the policy debate, which require platforms to feature at least two sides (or two groups, using our terminology).³ We believe that the broader definition serves the purposes of analyzing platform strategies (and, in particular, platform dynamics) better. The reason is that

² For a comprehensive textbook on this topic, see Fujita and Thisse (2013).

³ For instance, the European Commission (2015, p. 5) gave the following definition: ‘‘Online platform’’ refers to an undertaking operating in two (or multi)-sided markets, which uses the Internet to enable interactions between two or more distinct but interdependent groups of users so as to generate value for at least one of the groups’. See also Monopolkommission (2015) and House of Lords (2016).





the property to feature only within-group external effects within a single group or also cross-group external effects with a second group is in some cases endogenous, i.e., it is a decision by the intermediary, as we illustrate below. Of course, there is no right or wrong definition. However, we believe that a broader definition is more appropriate, as it allows us to apply some economic mechanisms to this broad set of phenomena. Later on, it will be useful to dedicate particular attention to two-sided platforms, where additional issues arise.

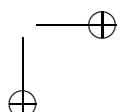
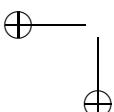
2.2 Illustrations

Amazon and Facebook were initially one-sided platforms. Even today, their attractiveness depends to a large extent on the strength of positive within-group external effects. Amazon, by establishing its Marketplace, added another group, namely independent sellers, which generates positive cross-group external effects between buyers and sellers. However, their main strength vis-à-vis rivals comes, arguably, from the positive within-group external effects on the buyer side, arising from its recommender and reputation systems (where the latter refers to the reviews and grades about products, not sellers). Facebook built its business model on providing advertisers the possibility of (targeted) advertising. To the extent that private users rather dislike advertising, Facebook is exploiting the user base it maintains because of the social networking benefits (which are within-group external effects) to provide benefits to advertisers. It then charges advertisers for this service. Here, users exert a positive cross-group external effect on advertisers, while advertisers exert a negative cross-group external effect on users. Clearly, advertising here serves as a monetization device (which benefits from a large, interconnected user base, as it allows for better targeting). However, the strength of Facebook in the market place is arguably due to within-group external effects arising from social networking among users. An alternative strategy by Facebook could have been to charge users for participation or usage. Using the narrower definition proposed elsewhere, Facebook would not be classified as a platform, even though the interaction between users would continue to be present.

Telecommunication networks also provide a nice illustration of the difficulty in drawing a clear line between within- and cross-group external effects. Most of the economic literature on telecommunication networks assumes, for simplicity, uniform calling patterns, i.e., an equal likelihood for each subscriber to call and be called by any other subscriber;⁴ this assumption of fully symmetric participants implied a single group exhibiting within-group external effects. Another simplifying assumption would be to consider that some people only make calls, while others only receive calls (e.g., restaurants and customers who want to order for delivery or make a reservation); in that case, there would be two distinct groups, with only cross-group external effects. The reality is naturally somewhere between these two extremes: subscribers are heterogeneous in their propensity both to make calls and to receive calls. Moreover, calling patterns are largely reported to be non-uniform: as indicated above, most subscribers have a ‘calling circle’, i.e., a subset of subscribers with whom they interact more frequently than with others.⁵ Seen from an individual subscriber’s perspective, external effects are then mostly within-group (i.e., inside the calling circle), cross-group effects (i.e., outside the calling circle)

⁴ See, for instance, Armstrong (1998), Laffont, Rey, and Tirole (1998a, 1998b), and De Bijl and Peitz (2002).

⁵ See, e.g., Hoernig, Inderst, and Valletti (2014).



being relatively limited; yet, each subscriber makes a different distinction between the two types of external effects as calling circles differ.

Finally, even if a platform facilitates the interaction between two distinct groups of users, some network effects may be jointly generated by all users, irrespective of the group they belong to, which may further blur the distinction between cross- and within-group external effects. Take the example of peer-to-peer marketplaces like Uber or Airbnb, which enable the interaction between providers and consumers of services; clearly, each group exerts positive cross-group external effects on the other group. Yet, the quality of the matching between peers from the two groups increases with the volume and reliability of data that the platforms collect from providers and consumers alike. Hence, a form of within-group external effects appears: the larger the participation on both sides, the more data is generated (about feedbacks, reputation, reviews, geo-localization, etc.), which enhances the quality of the platform's service and, thereby, the utility of *all* users.

3 ECONOMICS OF MARKETS WITH NETWORK EFFECTS

In this section, we explore the economics of markets with network effects. We first focus on the demand side and derive the demand for a good that exhibits network effects. We then turn to the supply side, considering first markets with a single network good, provided either by a monopoly or by perfectly competitive firms. We move next to markets with several network goods supplied by distinct firms, where decisions about prices and capacities cannot be separated from decisions about compatibility.

3.1 Demand for a Network Good

To determine user demand in a market with network effects, we consider an intermediation market with a large number of potential users.⁶ The total number of potential users has measure \bar{n} . Each user i is characterized by some stand-alone utility r_i of using the services of the intermediary irrespective of the number of users. Users are distributed according to some cumulative distribution function G defined on an interval $[r, \bar{r}]$. The cumulative distribution function is continuous on its support and takes values $G(r) = 0$ and $G(\bar{r}) = 1$.

For the sake of simplicity, we focus on the participation decision of each user and do not look at the usage intensity. Each user i pays a price p_i for obtaining the possibility to interact with other users via an intermediary. When interacting, we assume that each user obtains an additional utility $u_i(n)$ that depends on the measure of participating users n . We set $u_i(0) = 0$ for all users. In particular, if everybody with a stand-alone utility of r_i and higher participates we have $n = (1 - G(r_i))$. A user may decide not to participate and obtain an outside valuation of $v_0 = 0$ or to be active and obtain valuation $r_i + u_i(n) - p_i$.

We consider two alternative specifications of heterogeneity among users, one with respect to r_i and one with respect to u_i . Suppose first that all users have the same function $u(\cdot) = u_i(\cdot)$ and face the same price $p = p_i$. Consumers with a large r_i tend to participate, while those with a low r_i tend to be more reluctant. Suppose that there is an interior user with $\hat{r} \in (r, \bar{r})$ that satisfies $\hat{r} + u((1 - G(\hat{r}))) - p = 0$. Thus, when u is downward sloping or not too strongly

⁶ Our exposition on the demand for network goods follows Belleflamme and Peitz (2015, Chapter 20).

upward sloping, there exists a stable user equilibrium such that all consumers with $r_i \geq \hat{r}$ participate, while all users with $r_i < \hat{r}$ do not, with $\hat{r} + \varepsilon + u((1 - G(\hat{r} + \varepsilon))) - p > 0$ and $\hat{r} - \varepsilon + u((1 - G(\hat{r} - \varepsilon))) - p < 0$. There is a stable user equilibrium such that *all* users participate if the slope of u is positive and sufficiently large, and $\underline{r} - p$ not too small such that $\underline{r} + u(\bar{n}) - p \geq 0$.

It is important to note that in the presence of positive network effects, multiple equilibria may arise. That is, the same price may give rise to several equilibrium network sizes. For instance, if $\bar{r} < p = \hat{r} + u((1 - G(\hat{r}))) < \underline{r} + u(\bar{n})$, with $\hat{r} \in (\underline{r}, \bar{r})$, then on top of the interior equilibrium with a mass of $0 < 1 - G(\hat{r}) < \bar{n}$ users, no participation and full participation are also equilibria. As an illustration, suppose that r_i is uniformly distributed on $[0, 1]$, there is a mass one of users, and $u(n) = 2n$. Then, for $p = 3/2$, the following three situations are equilibria: (i) no user participates (as the user with the largest stand-alone utility is not willing to pay the price when no other user participates: $1 + 2 \times 0 < p = 3/2$); (ii) the set of users with $r_i \geq 1/2$ participates (as $r_i = 1/2$ is the indifferent user when the mass of participants is equal to $1/2$: $1/2 + 2 \times 1/2 = p = 3/2$); (iii) all users participate (as the user with the lowest stand-alone utility is willing to pay the price when all other users participate: $0 + 2 \times 1 > p = 3/2$).

A special case is the situation in which users are homogeneous and the stand-alone utility is set equal to zero. Then, if users suffer from the presence of more users with the intermediary, there will not be any interaction for any positive price. However, if users benefit from the presence of other users such that u is strictly increasing, there is a stable consumer equilibrium such that all consumers participate if $u(\bar{n}) > p$.

Suppose next that all users have the same stand-alone utility but differ in their valuation of the interaction with other users. That is, $r_i = r$ for all users, while $u_i(n) \neq u_j(n)$ for $i \neq j$. As there are multiple ways in which the functions $u_i(\cdot)$ may differ, let us develop one specific example to show that multiple equilibria may arise as well under this alternative specification of user heterogeneity. Suppose that there is a unit mass of users ($\bar{n} = 1$), identified by a parameter θ that is uniformly distributed on $[0, 1]$, and let user θ value the interaction with other users according to the function $u_\theta(n) = \theta n$. That is, we assume that network benefits increase linearly with the size of the network of participants but with a different intensity for each user. If all users face the same price p to interact through the intermediary, the indifferent user is identified as $\hat{\theta}$ such that $r + \hat{\theta}n = p$. As all users with $\theta \geq \hat{\theta}$ will participate, the mass of participants is equal to $n = 1 - \hat{\theta}$, which implies that $\hat{\theta} = 1 - n$. It follows that the inverse demand for participation can be written as $p = r + n(1 - n)$, which is a bell-shaped function of n that reaches a maximum at $n = 1/2$, where $p = r + 1/4$. So, for any price $p \in (r, r + 1/4)$, there are two ‘demand levels’, i.e., the two values of n that solve $p = r + n(1 - n)$. Moreover, $n = 0$ is also compatible with such a price as in this case (i.e., if no user participate), each user has a negative net utility and thus finds it optimal not to participate ($r + \theta \times 0 - p < v_0 = 0$ for all θ). As above, we can thus find prices for which three equilibrium participation levels coexist. They all stem from ‘*self-fulfilling prophecies*’ insofar as they correspond to network sizes that generate utilities such that the combined participation decisions of the users exactly generate these network sizes.

3.2 Monopoly Provision of a Network Good

Suppose that the network good is provided by a monopoly intermediary and that this intermediary is able to choose how many users to connect to the network. Suppose also that

there is a constant marginal cost, $c \geq 0$, to connect an extra user to the network. By choosing the size of the network, the intermediary has the potential to internalize network effects as it recognizes that a larger network raises the users' willingness to pay and, thereby, its revenues. We may then wonder whether a monopoly intermediary does not have the incentive to extend the network up to the size that would prevail under perfect competition. However, this is generally not the case, as we now show in a simple example.

Consider a pure communication technology that generates only network benefits (the stand-alone utility is equal to zero for all users). As above, suppose that there is a unit mass of users ($\bar{n} = 1$), identified by a parameter θ that is uniformly distributed on $[0, 1]$; user θ values the possibility to communicate with a mass n of other users according to the function $u_\theta(n) = \theta n$. Following the methodology developed above, we compute the inverse demand for network participation as $p = n(1 - n)$. Note that the maximum price is reached for $n = 1/2$ and is equal to $1/4$; we therefore assume that $c < 1/4$ to make the problem non-trivial. The intermediary chooses n to maximize $n^2(1 - n) - nc$. From the first- and second-order conditions, we find the profit-maximizing network size as $n^m = (1 + \sqrt{1 - 3c})/3$.

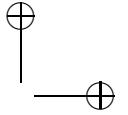
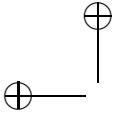
If the network good was supplied by a competitive industry, the network size n^c , would be such that $p = c$; that is, $n^c(1 - n^c) = c$, which is equivalent to $n^c = (1 + \sqrt{1 - 4c})/2$. It is easily checked that $n^c > n^m$, meaning that despite its ability to internalize network effects (which competitive firms lack), the monopoly intermediary still restricts the network size (i.e., the quantity) below the perfectly competitive level, thereby reducing welfare.

As far as welfare is concerned, it is important to note that in the presence of network effects, perfect competition generally also fails to achieve the first best. In our example, social welfare is maximized when all consumers join the network (i.e., $n^* = 1$) but, for any $c > 0$, $n^c < 1$.⁷ Network effects are the source of the market failure: when joining the network, users do not internalize the positive consumption externality that they exert on the other users.

3.3 Provision of Competing Network Goods

We now consider situations where users can choose among several substitutable network goods that are provided by different firms. As we will now discuss, the degree of compatibility among the various networks is of crucial importance since it conditions the network benefits that users enjoy when joining a particular network. If networks are fully incompatible, each firm makes up its own network, insofar as network benefits are product specific. At the other extreme, when networks are fully compatible, each user generates network benefits for any user of any network alike; it is as though a single network existed. In between, there are situations of imperfect compatibility where network benefits are stronger among users of the same network than among users of different networks. As an illustration, think of using your smartphone for its primary function, that is, to phone someone. If you use the regular phone lines, you do not need to worry about which operating system (OS) your correspondent has on her smartphone: there is full compatibility. In contrast, to make a video call through Facetime, users must both have an iPhone because this application, developed by Apple, is not available to users of devices running on other OSs than iOS; in this case, there is full incompatibility. As for partial compatibility, most 'Voice-over-IP' applications (e.g., Skype, WhatsApp or Viber)

⁷ Social welfare is computed as $W(n) = \int_{1-n}^1 \gamma n \, d\gamma - nc = \frac{1}{2}n^2(2 - n) - nc$. We check that for all $0 \leq n < 1$, $W(1) - W(n) = \frac{1}{2}(1 - n)(1 - 2c + n(1 - n)) > 0$.



are available for all major OSs; yet, the performance of some apps may vary across OSs, which may cause problems (such as synchronization issues) when calls are made between different devices.

3.3.1 A duopoly model

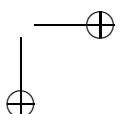
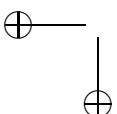
To analyze the importance of compatibility, we develop the following model of competition between two producers of network goods. We follow the seminal work of Katz and Shapiro (1985) and its extension by Crémer, Rey, and Tirole (2000).⁸ As for the demand side of the market, we assume that there are two types of users. On the one hand, a mass α of users have adopted the network good of firm A in the past and are now locked in by the contract they have previously signed. These consumers, who will be referred to as firm A 's ‘installed base’, make no decision in the game but confer an advantage to firm A over its competitor (we will therefore call firm A the ‘big firm’). On the other hand, there is a continuum of unattached users (of mass 1) who are identified by a parameter r , which is drawn from a uniform distribution on the unit interval. These unattached (or ‘new’) users choose whether to adopt the network good of firm A or of firm B . Letting r measure the stand-alone benefit of any network good, g_i measure the network benefit from good i , and p_i denote the price of good i , we have that a new user of type $r \in [0, 1]$ obtains a net surplus from adopting the good of firm i equal to $U_i(r) = r + g_i - p_i$. The network benefits generated by the two goods are defined respectively as $g_A = \nu(q_A + \alpha + \gamma q_B)$ and $g_B = \nu(q_B + \gamma q_A + \gamma \alpha)$. In these formulations, the parameter ν measures the strength of the network effects, q_i is the number of new users joining network i , and the parameter γ measures the degree of compatibility between the two goods. We assume that $\nu < 1/2$ to guarantee the existence of a stable equilibrium; as for γ , it is bounded below by zero (full incompatibility) and above by one (full compatibility).

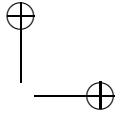
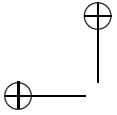
Regarding the supply side of the market, we assume that each good is produced at zero marginal cost. The two firms compete for the new users by choosing the capacity of their network (i.e., they compete à la Cournot). To derive the demand function facing each firm, we identify the new user r_0 who is exactly indifferent between the three options of joining network A , joining network B or not joining any: $r_0 + g_A - p_A = r_0 + g_B - p_B = 0$, which is equivalent to $r_0 = \hat{p}$, where $\hat{p} = p_A - g_A = p_B - g_B$ is the common ‘quality-adjusted price’ of the two firms. As all new users with $r \geq r_0$ decide to join one or the other network (and given our assumption of a uniform distribution of r), we have that the total number of users is given by $q_A + q_B = 1 - \hat{p}$. Using the definitions of \hat{p} , g_A and g_B , we can then write the inverse demand functions as

$$\begin{cases} p_A = 1 + \nu\alpha - (1 - \nu)q_A - (1 - \gamma\nu)q_B, \\ p_B = 1 + \gamma\nu\alpha - (1 - \nu)q_B - (1 - \gamma\nu)q_A. \end{cases}$$

We observe that the intercepts of both demands increase with the size of firm A 's installed base but to a lower extent for firm B than for firm A if compatibility is not complete ($\gamma < 1$);

⁸ Relatedly, Doganoglu and Wright (2006) analyze compatibility choice in a model with price-setting firms, where consumers can overcome incompatibility by multihoming – i.e., by buying the two incompatible versions. As they show, multihoming makes compatibility less attractive for firms, while it may make compatibility socially more desirable.





we also observe that incompatibility introduces a form of horizontal differentiation among the two network goods as the price of good i is more sensitive to a change in the capacity of network i (q_i) than of network j (q_j).

Firm i chooses its capacity q_i to maximize its profit $\Pi_i = p_i q_i$. Solving for the system of first-order conditions, one finds the following equilibrium quantities:

$$q_A^* = \frac{1}{3-(2+\gamma)\nu} + \frac{2-\gamma-\nu(2-\gamma)^2}{(3-(2+\gamma)\nu)(1-(2-\gamma)\nu)}\nu\alpha,$$

$$q_B^* = \frac{1}{3-(2+\gamma)\nu} - \frac{1-(2-\nu)\gamma}{(3-(2+\gamma)\nu)(1-(2-\gamma)\nu)}\nu\alpha.$$

From the first-order conditions, we easily find that $\Pi_i^* = (1 - \nu)(q_i^*)$.² As expected, the installed base α provides firm A with a competitive advantage as it allows it to reach a larger capacity than its rival at equilibrium: we check that

$$q_A^* - q_B^* = \frac{(1 - \gamma) \nu}{1 - (2 - \gamma) \nu} \alpha > 0.$$

We observe that the larger the α , the larger the advantage. This does not mean, however, that a sufficiently large installed base would necessarily drive firm B out of the market. If the degree of compatibility is large enough, more precisely, if $\gamma > 1/(2 - \nu)$, then firm B 's users benefit sufficiently from the network effects generated by firm A 's installed base to guarantee that $q_B^* > 0$.⁹ This also explains why enhanced compatibility reduces the difference between the equilibrium capacities of the two firms:

$$\frac{d(q_A^* - q_B^*)}{d\gamma} = -\frac{1 - \nu}{(1 - (2 - \gamma) \nu)^2} \alpha \nu < 0.$$

As compatibility increases, the perceived quality differential between the two network goods is reduced, which reduces firm A 's initial advantage.

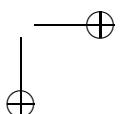
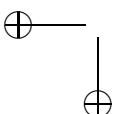
The latter finding suggests that the big firm may prefer more incompatibility. However, compatibility has an upside for both firms as it enhances the users' willingness to pay. It is indeed clear that users are better off when compatibility is improved. To see this, note that the consumer surplus is equal to $\frac{1}{2}(q_A^* + q_B^*)^2$ and that

$$q_A^* + q_B^* = \frac{2 + (1 + \gamma) \nu \alpha}{3 - (2 + \gamma) \nu}, \text{ with } \frac{d(q_A^* + q_B^*)}{d\gamma} = \frac{2 + (3 - \nu) \alpha}{(3 - (2 + \gamma) \nu)^2} \nu > 0.$$

3.3.2 Ex ante vs ex post standardization

Collecting the previous results, it appears that the small firm (firm B here) unambiguously prefers more compatibility as it reduces its disadvantage with respect to the big firm and it expands demand. As for the big firm (firm A here), more compatibility is a mixed blessing: it expands demand but it attenuates the initial competitive advantage conferred by the installed base. As the latter effect depends on the size of the installed base, we expect the big firm to

⁹ For values of $\gamma < 1/(2 - \nu)$, firm B stays on the market provided that $\alpha \leq (1 - (2 - \gamma) \nu) / (\nu(1 - (2 - \gamma) \nu))$.



prefer full compatibility to full incompatibility if the installed base is not too large. We check indeed that

$$\Pi_A^*|_{\gamma=1} \geq \Pi_A^*|_{\gamma=0} \Leftrightarrow \alpha \leq \frac{1 - 2\nu}{3 - 4\nu + 2\nu^2} \equiv \bar{\alpha}.$$

Suppose now for simplicity that compatibility is all or nothing and that it has to be agreed upon by the two firms to be achieved (i.e., it is technically or legally impossible to achieve compatibility on a unilateral basis). Then, the previous analysis teaches us that if the installed base of the big firm is not too large (i.e., if $\alpha \leq \bar{\alpha}$), both firms will prefer compatibility. For instance, they will agree to adopt the same specifications for their network goods so as to make them fully interoperable. In that case, *ex ante* (or ‘de jure’) standardization prevails. Typically, this form of standardization follows from negotiations among firms that take place within standard-setting organizations (SSOs).¹⁰

In contrast, if the big firm starts the game with a significant installed base (i.e., if $\alpha > \bar{\alpha}$), then it will not agree with the small firm to achieve compatibility. Both firms will then push their own specification and let users choose between two incompatible network goods. A so-called ‘standards war’ will ensue, with both firms fighting to become the *ex post* (or ‘de facto’) standard (i.e., the specification that will eventually gain widespread acceptance).¹¹

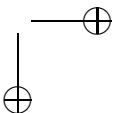
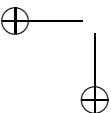
3.3.3 Strategies in standards wars

As we have just seen, the existence of a larger installed base for a firm not only confers a competitive advantage to this firm but may also lead it to prefer incompatibility. If incompatibility is the chosen course of action, one understands then that firms may have an incentive to build an installed base today so as to increase their chances to win the standards war tomorrow. They may thus want to set lower prices today (so as to attract more users) with the hope of being able to set higher prices tomorrow (as users will be willing to pay more to join a larger network).

To analyze this kind of issue, we need to move away from the static framework that we have used so far and adopt instead a dynamic approach, where firms and users make decisions over consecutive periods. Cabral (2011) develops a model of dynamic competition between firms producing incompatible network goods. Users are assumed to live for potentially many periods (i.e., they die and are replaced with a constant hazard rate). As above, users derive stand-alone and network benefits from the network good that they choose. The stand-alone component is received once the user joins a network and its value is supposed to be the user’s private information; the network component is received each period that a consumer is still alive. A newborn user chooses one of the two existing networks and stays with it until death. This decision is assumed to be made in a rational, forward-looking way. This means that users are able to anticipate all future decisions so as to estimate correctly the evolution of network

¹⁰ There are a very large number of SSOs, working at different levels: international (e.g., the International Telecommunication Union, ITU), regional (e.g., the European Committee for Standardization, CEN), or sectorial (e.g., the World Wide Web Consortium, W3C).

¹¹ Well-known standards wars in the consumer electronics sector are the wars of formats for videotape recording (with VHS winning over Betamax) and for high-definition optical discs (with Blu-ray discs supplanting HD DVD).



sizes. In each period, the two firms compete for new consumers to join their network by setting the price of their network good (prices below marginal costs are allowed).

In this framework, a firm with a larger network faces an interesting trade-off when setting its prices: taking a short-term perspective, the big firm should set higher prices as it is more attractive to users ('harvesting' effect); however, when taking future payoffs into account, the firm has an incentive to set lower prices because the gains from increasing network size are larger for a big than for a small firm ('investment' effect). It is *a priori* not clear which of these effects will dominate, meaning that price functions may be increasing or decreasing in time.

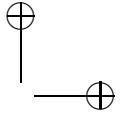
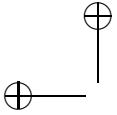
Although the pricing equilibrium is symmetric, market shares are generally asymmetric because of the stochastic appearance of new users. There are then two questions of interest: (i) Does the large network attract a new user with higher probability than the small network? (ii) Does the large network increase its size in expected value? The answer to the first question is yes. As for the second question, the answer is yes as well, as long as network effects are sufficiently strong and the big firm is still shy of holding 100 percent of the market.

The previous results suggest that getting a headstart may prove valuable in a standards wars. However, the model abstracts away a number of factors that may reverse this statement. First, both network goods are assumed to have the same, fixed, intrinsic quality (which is measured by the value that users attach to the stand-alone benefits). This is a simplifying assumption as product qualities result from firms' investments in R&D and are likely to improve over time (because of knowledge spillovers and/or users' feedback). Firms may then face the following trade-off when deciding to enter early on the market: they may secure a headstart but they may also have to compromise on the quality of their product; a later entrant with a better product would then be able to overcome an initial disadvantage.

A second issue concerns compatibility across periods: by posing that users can collect network benefits during their whole life, the model implicitly assumes that new versions of the network goods are backward compatible with old ones. In reality, firms may decide against backward compatibility. One reason is that firms may want to force old users to buy a new version of the network good (because, if they stick to the old version, they will not enjoy network benefits from users of the new, incompatible, version). This form of 'planned obsolescence' is commonplace in consumer electronics markets.¹² Another reason, which is hard to disentangle from the previous one, is that backward compatibility often constrains firms in their efforts to improve their products; to reach the full potential of technological advancement, they may thus decide to introduce a new version that is incompatible with the previous one. Abandoning backward compatibility clearly modifies the incentives to build an early installed base of users.

Finally, when users are not as fully rational as they are supposed to be in the previous framework, there may be less costly ways to get a headstart than through low introductory prices. If users form their expectations in a non-rational way, firms have the potential to influence these expectations in their favor, thereby creating self-fulfilling prophecies: if users are made to believe that a particular firm will win the standards war, then they will adopt the product of that firm, thereby helping it to win the war and confirming the initial beliefs.

¹² Choi (1994) examines this issue in a monopoly framework.



4 MANAGING NETWORK EFFECTS ON TWO-SIDED PLATFORMS

In this section, we turn to platform markets, where intermediaries facilitate the interaction between separate groups of users. To understand how cross-group external effects shape the strategies of intermediaries, we focus first on the decisions of a single intermediary. We then consider competition among intermediaries in the presence of positive cross-group external effects; we do so in two distinct settings: when markets tip (i.e., when all users aggregate on a single platform at equilibrium) and when they do not. We next examine media markets, where one group of users may exert negative cross-group effects on the other. Finally, we explore a number of issues that were abstracted away in the previous analyses, namely advanced pricing strategies, within-group external effects, and investment issues.

4.1 Monopoly Pricing by a Two-sided Platform

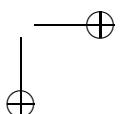
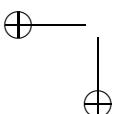
Consider a platform that invites two groups of agents, e.g., sellers and buyers, to interact via the platform. Suppose that the platform charges a membership fee m_s on the seller side and a membership fee of m_b on the buyer side. Thus, the platform's profits are $\Pi = n_s(m_s - c_s) + n_b(m_b - c_b)$, where n_k is the number of participants on side $k \in \{s, b\}$ and c_k is the cost per participant on the respective side. In a two-sided market, participation on at least one side of the market depends on participation on the other side. Thus, in general, demand on one side depends on prices on both sides of the market.

Participants on side i obtain a stand-alone utility of r_k (which corresponds to the intrinsic willingness to pay, when participation on the other side is assumed to be nil) plus some utility that depends on the number of participants on the other side. It may also depend on the participation level on the own side (we will address this possibility in Section 4.5.2). Let us postulate a positive cross-group external effect: a participant on side s benefits from more participation on side b and vice versa. For the sake of simplicity, let us, for the moment, assume that this relationship is linear. Thus the gross utility of a participant on side s is $r_s + \pi n_b$ and of a participant on side b is $r_b + u n_s$, where π and u are the marginal external effect enjoyed by sellers and buyers, respectively.

Participants have heterogeneous outside options; for simplicity, we assume that there is a unit mass on each interval of length 1 in the range $[0, V]$, where V is sufficiently large such that some participants on each side do not participate in the solutions we are going to consider. Hence, all those participants on side s with an outside option of less than $v_s = r_s + \pi n_b - m_s$ will pay the membership fee if they expect a participation level of n_b , and all those participants on side b with an outside option of less than $v_b = r_b + u n_s - m_b$ will pay the membership fee if they expect a participation level of n_s . Hence, since buyers and sellers are uniformly distributed as specified above, we have $n_s = v_s$ and $n_b = v_b$. For given membership fees, participants play an anonymous game and we solve for the Nash equilibrium of this game; the expected number of participants on each side has to be equal to the actual number. Hence, we solve the system of two linear equations in two variables, n_s and n_b and obtain

$$n_s = \frac{r_s + \pi s_b - m_s - \pi m_b}{1 - \pi u} \text{ and } n_b = \frac{r_b + u s_s - m_b - u m_s}{1 - \pi u}.$$

We assume that $\pi u < 1$ (i.e., cross-group external effects are not too strong), so that the numbers of agents registering on the platform are decreasing functions of the membership fees.



We can now solve the platform's maximization problem. The first-order conditions with respect to m_s and m_b can be written respectively as

$$\begin{cases} m_s = \frac{1}{2}(c_s + r_s) - \frac{1}{2}(\pi + u)m_b + \frac{1}{2}(\pi r_b + uc_b), \\ m_b = \frac{1}{2}(c_b + r_b) - \frac{1}{2}(\pi + u)m_s + \frac{1}{2}(ur_s + \pi c_s). \end{cases}$$

We observe that the presence of positive cross-group effects affects the platform's choice of fees in two ways. First, positive cross-group effects generate a negative relationship between the two fees; because participation on the two sides are complementary to one another, lowering the fee on one side drives the platform to raise the fee on the other side.¹³ Second, the cross-group effects make the optimal fee on one side also depend on features (cost, willingness to pay) pertaining to the other side. This is another consequence of the complementarity between the two sides: the opportunity cost of attracting, say, an additional buyer is lower than the marginal cost (c_b) because this additional buyer will entice extra participation on the seller side and hence, extra revenues and costs for the platform on that side (which depend on r_s and c_s). In general, because of cross-group effects, revenues and costs cannot be easily allocated to one side or the other.

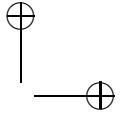
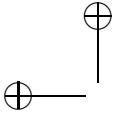
We can now proceed by solving the previous system to obtain the optimal fees.¹⁴ To clarify the intuition, we introduce the following notation: let $\mu_k \equiv r_k - c_k$ denote the difference between the intrinsic willingness to pay (r_k) and the marginal cost (c_k) on side k ($k = s, b$). We can then write:

$$\begin{cases} m_s^* - c_s = \frac{1}{2}\mu_s + \frac{1}{2}\frac{\pi-u}{4-(u+\pi)^2}(2\mu_b + (u+\pi)\mu_s), \\ m_b^* - c_b = \frac{1}{2}\mu_b + \frac{1}{2}\frac{u-\pi}{4-(u+\pi)^2}(2\mu_s + (u+\pi)\mu_b). \end{cases}$$

We have expressed the optimal margins (i.e., the optimal fee minus the marginal cost) as the sum of two terms. The first term is the margin that platforms would set absent cross-group effects (i.e., for $\pi = u = 0$); in that case, the margin on each side would only depend on costs and willingness to pay on that side (it is as though the 'platform' was selling two independent services to two separate groups of customers). The second term depends on the intensity of the cross-group effects and on parameters pertaining to both sides. Interestingly, in this linear model, this second term vanishes in the special case where the marginal cross-group external effects are equal across sides ($\pi = u$). In contrast, if $\pi > u$, i.e., if sellers value the interaction with buyers more than buyers value the interaction with sellers, we see that the platform chooses to have a larger margin on the seller side and a lower margin on the buyer side (with respect to what would prevail in the absence of cross-group effects). The intuition is simple: as the two sides are complementary, the platform can attract more agents on one side by lowering its fee on the other side; when, for instance, $\pi > u$, this 'leverage strategy'

¹³ In that regard, two-sided platforms bear some resemblance to multiproduct firms. Yet, as Rochet and Tirole (2003) point out, end users internalize the corresponding externalities in a multiproduct setting but not in a multisided setting.

¹⁴ To satisfy the second-order conditions for profit maximization, we need to impose a more stringent condition than $e_s e_b < 1$, namely $(e_s + e_b)^2 < 4$.



is more effective when applied on the buyer side; it is thus profitable for the platform to lower the buyer fee and to capture the extra value created on the seller side by increasing the seller fee. This reasoning may even lead the platform to set a negative margin on the buyer side: it can be checked that for $\pi > u$ and $\mu_b/\mu_s < (\pi - u)/(2 - \pi(\pi + u))$, we have $m_b^* < c_b$. The exact reverse logic applies in the case where $u > \pi$.

Finally, we compute the platform's profit at the optimum as

$$\Pi^* = \frac{\mu_s^2 + \mu_b^2 + \mu_s \mu_b (\pi + u)}{4 - (\pi + u)^2}.$$

We see clearly in the latter expression that the platform's profit increases with the intensity of the cross-group external effects (i.e., u and π).

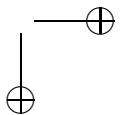
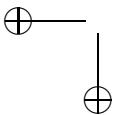
More generally, if the distribution of agents' types is not linear, we obtain increasing and monotone relationships $n_s = N_s(v_s)$ and $n_b = N_b(v_b)$, respectively. As illustrated above for the linear case, the platform maximizes $N_s(v_s)(m_s - c_s) + N_b(v_b)(m_b - c_b)$ with respect to m_s and m_b , where the participation levels $N_s(v_s)$ and $N_b(v_b)$ depend on membership fees.

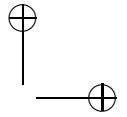
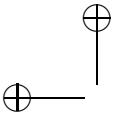
4.2 Pricing Under Platform Competition When Markets Tip

We now examine the competition between two platforms, which we call 1 and 2. We continue to consider two groups of agents, which we continue to call, for convenience, sellers (group s) and buyers (group b). We focus here on situations where only one platform survives at equilibrium; it is then said that the market 'tips' in the sense that all agents end up interacting on a single platform. Ingredients for such a result are positive and strong cross-group effects, closely substitutable platforms, and single homing.

In the model we develop, adapted from Caillaud and Jullien (2003), the platforms offer exactly the same services and are thus seen, other things being equal, as perfect substitutes by the agents. Both groups, sellers and buyers, are supposed to consist of a continuum of mass 1. There exist positive cross-group external effects between the two groups insofar as each agent uses the matching services of one or the other platform to find the unique trading partner she has in the other group. Hence, the probability of finding one's partner on a particular platform increases with the number of agents of the other group that register with this platform. In particular, if n_s^i sellers (resp. n_b^i buyers) register with platform i ($i = 1, 2$), then the probability for a buyer (resp. a seller) to find her match on platform i is equal to λn_s^i (resp. λn_b^i) where λ is the probability that two matching partners find each other when they register with the same platform. The gross gain from a successful match is equal to 1/2 for each agent (gains of trade are normalized to 1 and are supposed to be equally shared among trading partners after some efficient bargaining process). The net gain is then $1/2(1 - p^i)$, where p^i is the transaction fee that platform i charges.¹⁵ Given that platforms also set membership fees m_k^i , the expected utilities for a seller and for a buyer when registering with platform i along with n_s^i sellers and n_b^i buyers are respectively equal to $U_s^i = \lambda n_b^i \frac{1}{2}(1 - p^i) - m_s^i$ and $U_b^i = \lambda n_s^i \frac{1}{2}(1 - p^i) - m_b^i$.

¹⁵ Because we assume constant gains from trade and efficient bargaining, transaction fees are non-distortionary. As a result, it is only the total fee that matters (i.e., p^i) and not the way it is split between the trading partners.





We analyze the following game. In the first stage, platforms set their price structure to maximize their profit; that is, platform i chooses the triple (m_s^i, m_b^i, p^i) to maximize $\Pi^i = n_s^i(m_s^i - c_s) + n_b^i(m_b^i - c_b) + \lambda n_s^i n_b^i p^i$, where c_k ($k = s, b$) is the constant cost a platform incurs when providing services to one agent of type k (with $c_s + c_b < \lambda$ so that total gains from trade are larger than total costs). In the second stage, agents choose which platform (if any) to register with; we assume single homing on both sides (agents register with at most one platform) and we normalize the outside option to zero.

In this Bertrand competition, platforms resort to a ‘divide-and-conquer’ strategy, which consists in first ‘dividing’ by subsidizing one group of agents to convince them to join and next, in ‘conquering’ the agents of the other group, who have no better option than to join the platform as well. To be sure of attracting the group of, say, buyers, platform i must offer them a better deal than platform j , even in the worst-case scenario where buyers hold the pessimistic belief that they will find no seller on platform i ; that is, it must be that $-m_b^i > \lambda \frac{1}{2}(1 - p^j) - m_b^j$. If this condition is met, buyers will join platform i and sellers will follow suit (whatever their beliefs), thereby generating maximal aggregate surplus $\lambda - c_s - c_b$, which platform i can capture by setting the transaction fee at its maximal level ($p^i = 1$). Platform i finds this strategy profitable as long as $\lambda - c_s - c_b + m_s^i \geq -m_b^i$. Yet, as platform j can act in exactly the same way, competition through divide-and-conquer strategies will drive profits to zero and allow only one platform to remain active. At equilibrium, the remaining platform subsidizes full participation, charges the maximal transaction fee ($p^i = 1$) and makes zero profit ($m_s^i + m_b^i = c_s + c_b - \lambda$). As the presence of positive cross-side external effects makes it efficient to have all agents register with the same platform, the equilibrium is socially desirable.

If one platform (the incumbent) could play before the other (the entrant), the same equilibrium would prevail, with the incumbent deterring the entry but foregoing all profit. The surviving platforms may, however, achieve positive profits at equilibrium when transaction fees are not feasible (e.g., because agents, when matched, could bypass the platform to trade).¹⁶

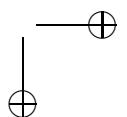
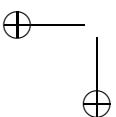
4.3 Pricing Under Platform Competition When Markets Do Not Tip

We now introduce some differentiation between the platforms, so as to add a ‘dispersion’ force that can counterbalance the ‘agglomeration’ force exerted by the combination of the positive cross-group effects. We also explicitly consider the possibility for the members of one group to be active on two platforms at the same time (so-called multihoming). To this end, we follow the approach proposed by Armstrong (2006) and Armstrong and Wright (2007).

4.3.1 Two-sided single homing

The basic ingredients are the same as in the model with two competing matching intermediaries: a unit mass of sellers and a unit mass of buyers having to choose at most one platform on which they will interact, with this interaction generating positive cross-group external effects. In contrast with the previous model, we assume now that platforms only compete in membership fees and that agents perceive them as horizontally differentiated. Horizontal differentiation is modeled in the Hotelling fashion: platforms are located at the extreme points

¹⁶ See Caillaud and Jullien (2001, 2003).



of the unit interval; sellers and buyers are uniformly distributed on this unit interval and incur an opportunity cost of visiting a platform that increases linearly in distance at rates τ_s and τ_b , respectively; participation is sufficiently attractive to drive all buyers and sellers to be active on one platform (it follows that on each side, the total number of agents on the two platforms adds up to 1: $n_s^1 + n_s^2 = n_b^1 + n_b^2 = 1$).

The nature of the interaction on the platform is the following: sellers offer perfectly differentiated products and buyers purchase one unit from each seller active on the platform; each seller makes a profit per buyer of π and each buyer derives utility u per seller.¹⁷ Hence, seller and buyer surpluses gross of any opportunity cost of visiting platform i are

$$v_s^i = r_s + n_b^i \pi - m_s^i \quad \text{and} \quad v_b^i = r_b + n_s^i u - m_b^i,$$

where m_b^i and m_s^i are the membership fees set by platform i , and r_b and r_s are the stand-alone benefits. The seller and the buyer who are indifferent between the two platforms are respectively located at x_s and x_b such that $v_s^1 - \tau_s x_s = v_s^2 - \tau_s (1 - x_s)$ and $v_b^1 - \tau_b x_b = v_b^2 - \tau_b (1 - x_b)$. It follows that $n_s^1 = x_s$, $n_s^2 = 1 - x_s$, $n_b^1 = x_b$, and $n_b^2 = 1 - x_b$. Combining these equations together with the expressions of v_s^i and v_b^i , we obtain the following expressions for the numbers of buyers and sellers at the two platforms:

$$\begin{cases} n_s^i(n_b^i) = \frac{1}{2} + \frac{1}{2\tau_s} \left((2n_b^i - 1)\pi - (m_s^i - m_s^j) \right), \\ n_b^i(n_s^i) = \frac{1}{2} + \frac{1}{2\tau_b} \left((2n_s^i - 1)u - (m_b^i - m_b^j) \right). \end{cases} \quad (11.1)$$

Solving the previous system, we derive the number of buyers and sellers as a function of the membership fees:

$$n_s^i(m_s^i, m_s^j, m_b^i, m_b^j) = \frac{1}{2} + \frac{\pi(m_b^j - m_b^i) + \tau_b(m_s^j - m_s^i)}{2(\tau_b\tau_s - u\pi)},$$

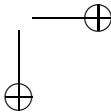
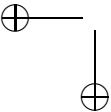
$$n_b^i(m_s^i, m_s^j, m_b^i, m_b^j) = \frac{1}{2} + \frac{u(m_s^j - m_s^i) + \tau_s(m_b^j - m_b^i)}{2(\tau_b\tau_s - u\pi)},$$

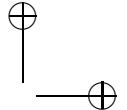
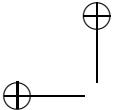
where we assume that $\tau_b\tau_s > u\pi$, i.e., that the transportation cost parameters τ_b and τ_s (which measure the perceived horizontal differentiation between the two platforms) are sufficiently large with respect to the gains from trade u and π (which measure the cross-group external effects). Under this assumption, the number of members of one group at one platform decreases not only with the membership fee that they have to pay but also with the membership fee that the other group has to pay on this platform.¹⁸

Platform i chooses m_s^i and m_b^i to maximize $\Pi^i = (m_s^i - c_s)n_s^i(\cdot) + (m_b^i - c_b)n_b^i(\cdot)$, where we assume as before that platforms incur costs c_s and c_b when they register additional sellers

¹⁷ It is assumed, quite realistically, that in a seller-buyer relationship, prices or terms of transactions are independent of the membership fee that applies to buyers and sellers.

¹⁸ For stronger cross-group external effects and/or weaker horizontal differentiation (i.e., for $u\pi > \tau_b\tau_s$), the number of agents on one platform would be an *increasing* function of their membership fee. The market would then naturally tip as in the model of the previous section.





and buyers. At the symmetric equilibrium ($m_s^1 = m_s^2 \equiv m_s$ and $m_b^1 = m_b^2 \equiv m_b$), the first-order conditions can be written as¹⁹

$$\begin{cases} m_s = c_s + \tau_s - \frac{u}{\tau_b}(\pi + m_b - c_b), \\ m_b = c_b + \tau_b - \frac{\pi}{\tau_s}(u + m_s - c_s). \end{cases}$$

The equilibrium membership fee for the sellers is equal to marginal costs plus the product-differentiation term as in the standard Hotelling model, adjusted downward by the term $\frac{u}{\tau_b}(\pi + m_b - c_b)$. To understand this term, note from expression (11.1) that each additional seller attracts u/τ_b additional buyers. These additional buyers allow the intermediary to extract π per seller without affecting the sellers' surplus. In addition, each of the additional u/τ_b buyers generates a margin of $m_b - c_b$ to the platform. Thus $\frac{u}{\tau_b}(\pi + m_b - c_b)$ represents the value of an additional buyer to the platform. The same holds on the buyers' side.

Solving the system of first-order conditions gives explicit expressions for equilibrium membership fees and platforms' profits:²⁰

$$m_s^* = c_s + \tau_s - u \text{ and } m_b^* = c_b + \tau_b - \pi,$$

$$\Pi^{1*} = \Pi^{2*} = \frac{1}{2}(\tau_b + \tau_s - u - \pi).$$

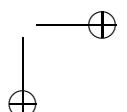
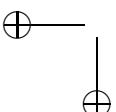
We observe that the equilibrium membership fee for one group is equal to the usual Hotelling formulation (marginal cost plus transportation cost) adjusted downward by the cross-group external effect that this group exerts on the other group. As for the platforms' equilibrium profits, they are increasing in the degree of product differentiation on both sides of the market (as in the Hotelling model) and decreasing in the buyers' and sellers' surplus for each transaction, i.e., the magnitude of the cross-group external effects. The intuition for the latter result is the following: as cross-group external effects increase, platforms compete more fiercely to attract additional agents on each side as they become more valuable.

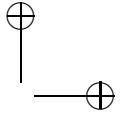
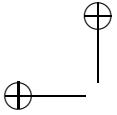
4.3.2 Multihoming on one side (competitive bottlenecks)

Suppose now that sellers have the possibility to multihome (i.e., to be active on both platforms at the same time), while buyers continue to single home. Sellers are then divided into three subintervals on the unit line: those sellers located 'on the left' register with platform 1 only, those located 'around the middle' register with both platforms, and those located 'on the right' register with platform 2 only. At the boundaries between these intervals, we find the sellers who are indifferent between visiting platform 1 (resp. 2) and not visiting any platform; their locations are found as, respectively, x_{10} such that $r_s + n_b^1 \pi - m_s^1 = \tau_s x_{10}$, and x_{20} such that $r_s + n_b^2 \pi - m_s^2 = \tau_s (1 - x_{20})$. We assume for now that $0 < x_{20} < x_{10} < 1$ (we provide necessary and sufficient conditions below), so that $n_s^1 = x_{10}$ and $n_s^2 = 1 - x_{20}$, with the multihoming sellers being located between x_{20} and x_{10} . As far as buyers are concerned,

¹⁹ The second-order conditions require $4\tau_s \tau_b > (u + \pi)^2$, which is a more restrictive condition than $\tau_b \tau_s > u\pi$.

²⁰ We assume that $r_b + u/2 + \pi > c_b + 3/2\tau_b$ and $r_s + \pi/2 + u > c_s + 3/2\tau_s$ to guarantee full participation on the two sides.





we have the same situation as before. The number of buyers and sellers visiting each platform are thus respectively given by

$$n_b^i = \frac{1}{2} + \frac{u(n_s^i - n_s^j) - (m_b^i - m_b^j)}{2\tau_b} \text{ and } n_s^i = \frac{r_s + n_b^i\pi - m_s^i}{\tau_s}.$$

Solving this system of four equations in four unknowns, we get

$$\begin{cases} n_b^i = \frac{1}{2} + \frac{u(m_s^j - m_s^i) + \tau_s(m_b^j - m_b^i)}{2(\tau_b\tau_s - u\pi)}, \\ n_s^i = \frac{\pi}{\tau_s} \left(\frac{1}{2} + \frac{u(m_s^j - m_s^i) + \tau_s(m_b^j - m_b^i)}{2(\tau_b\tau_s - u\pi)} \right) + \frac{r_s - m_s^i}{\tau_s}. \end{cases}$$

The maximization problems of the two platforms are the same as above. Platform 1's best responses are implicitly defined by the first-order conditions, which can be expressed as

$$m_b^1 = \frac{-(u + \pi)m_s^1 + um_s^2 + \tau_s m_b^2 - \pi(u - c_s) + \tau_s(\tau_b + c_b)}{2\tau_s},$$

$$m_s^1 = \frac{-(u + \pi)\tau_s m_b^1 + u\pi m_s^2 + \pi\tau_s m_b^2 - \pi u(\pi + c_s + 2r_s) + u\tau_s c_b + (\pi + 2c_s + 2r_s)\tau_b\tau_s}{2(2\tau_b\tau_s - u\pi)}.$$

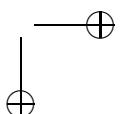
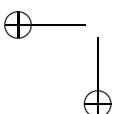
Second-order conditions require that $8\tau_b\tau_s > \pi^2 + u^2 + 6\pi u$. This condition is also sufficient to have a unique and stable interior equilibrium. Solving the previous system of equations, we find the equilibrium membership fees:

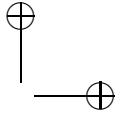
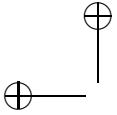
$$m_b^* \equiv m_b^{1*} = m_b^{2*} = c_b + \tau_b - \frac{\pi}{4\tau_s}(\pi + 3u + 2r_s - 2c_s),$$

$$m_s^* \equiv m_s^{1*} = m_s^{2*} = \frac{1}{2}(r_s + c_s) + \frac{1}{4}(\pi - u).$$

On the seller side, platforms have monopoly power. If the platform focused only on sellers, it would charge a monopoly price equal to $(r_s + c_s)/2 + \pi/4$ (assuming that each seller would have access to half of the buyers and, therefore, would have a gross willingness to pay equal to $\pi/2$). We observe that this price is adjusted downward by $u/4$ when the cross-group effect that sellers exert on the buyer side is taken into account. Similarly, on the buyer side, platforms charge the Hotelling price, $c_b + \tau_b$, less a term that depends on the size of the cross-group effects and on the parameters characterizing the seller side (r_s , c_s , and τ_s).

It is useful to compare price changes in the competitive bottleneck model to those in the two-sided single-homing model. In equilibrium, we observe that the membership fee for sellers is increasing in the strength of the cross-group effect ($\partial m_s^*/\partial\pi > 0$), whereas it is constant in the two-sided single-homing model. This is due to the monopoly pricing feature on the multihoming side. Everything else equal, if sellers are multihoming, the platform operators directly appropriate part of the rent generated on the multihoming side by setting higher membership fees. This is not the case in the single-homing world, where the membership fee





does not react to the strength of the network effect on the same side since platforms compete for sellers (and buyers).

It follows that at equilibrium,

$$n_b^{1*} = n_b^{2*} = \frac{1}{2} \text{ and } n_s^{1*} = n_s^{2*} = \frac{1}{4\tau_s} (u + \pi + 2(r_s - c_s)).$$

We still need to check under which conditions some (but not all) sellers multihome at equilibrium. This is so provided that $1/2 < n_s^{i*} < 1$, which is equivalent to $2\tau_s < \pi + u + 2(r_s - c_s) < 4\tau_s$. Under these conditions, the equilibrium net surplus of sellers and buyers (gross of transportation cost and for one platform) are equal to:

$$\begin{aligned} v_s^* &= \frac{1}{4}(u + \pi) + \frac{1}{2}(r_s - c_s), \\ v_b^* &= \frac{1}{4\tau_s}(u^2 + 4\pi u + \pi^2 + 2(u + \pi)(r_s - c_s)) + r_b - \tau_b - c_b. \end{aligned}$$

Note that v_s^* is the per platform seller's surplus.²¹ We observe that v_s^* and v_b^* are increasing in the net gain of the other side and in the net gain of the own side. The intermediaries' equilibrium profits are

$$\Pi^{i*} = \frac{1}{16\tau_s}(8\tau_b\tau_s - (\pi^2 + u^2 + 6\pi u) + 4(r_s - c_s)^2) > 0.$$

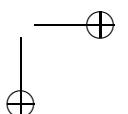
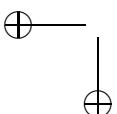
4.3.3 Single-homing vs multihoming environments

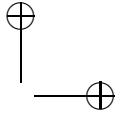
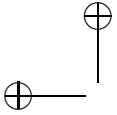
How do sellers and buyers surpluses compare in the two environments?²² In the model in which sellers multihome, platforms hold an exclusive access to their set of single-homing buyers (the ‘bottleneck’), which makes buyers valuable to extract profits on the seller side. We thus expect platforms to compete fiercely for buyers and, in return, to milk sellers. Hence, we may expect lower prices on the buyer side and higher prices on the seller side when compared to the two-sided single-homing model. We call this the ‘bottleneck effect’. However, this view can be challenged, as under multihoming with $n_s^{i*} > 1/2$, there are more sellers active on a platform than under single homing, thus adding value to participation on the buyer side. We call this the ‘expansion effect’. Moreover, multihoming sellers have access to all buyers (but pay twice the prices and the transportation costs).

As we illustrate next, whether buyers and sellers prefer one or the other environment depends on the parameter values. Using superscripts C and S to represent the competitive bottleneck and two-sided single-homing models, we can write the differences in surplus between the two environments as follows. For buyers, we have $v_b^C - v_b^S = (n_s^C - \frac{1}{2})u - (m_b^C - m_b^S)$; for single-homing sellers, we have $v_s^C - v_s^S = m_s^S - m_s^C$; as for multihoming sellers, we focus on the one located at the middle of the Hotelling line for whom the surplus difference is equal to $2v_s^C - \tau_s - (v_s^S - \frac{1}{2}\tau_s)$. Developing the latter expression, we find that this seller is better off in the competitive bottleneck environment than in the single-homing environment if and only if $\tau_s > u$.

²¹ Sellers located between $1 - n_s^{i*}$ and n_s^{i*} multihome and, therefore, earn a surplus of $2v_s^*$. On the other hand, v_s^* is the surplus earned by the sellers located between 0 and $1 - n_s^{i*}$, who choose to visit platform 1 only, and by the sellers located between n_s^{i*} and 1, who choose to visit platform 2 only.

²² This subsection closely follows Belleflamme and Peitz (2016).





Comparing prices, we see that the bottleneck effect dominates if

$$\begin{aligned} m_s^C > m_s^S &\Leftrightarrow \frac{1}{2}(r_s + c_s) + \frac{1}{4}(\pi - u) > c_s + \tau_s - u, \\ m_b^C < m_b^S &\Leftrightarrow c_b + \tau_b - \frac{\pi}{4\tau_s}(3u + \pi + 2r_s - 2c_s) < c_b + \tau_b - \pi. \end{aligned}$$

Simple computations show that the two conditions are equivalent and boil down to $K > 4\tau_s - 2u$, with $K \equiv \pi + u + 2(r_s - c_s)$. We recall from the analysis of the competitive bottleneck model that the following conditions are necessary for some (but not all) sellers to multihome: $2\tau_s < K < 4\tau_s$. Note that $4\tau_s - 2u < 2\tau_s$ whenever $\tau_s < u$.

We can thus distinguish between two cases. First, if $\tau_s < u$, it is always true that sellers pay higher and buyers lower fees in the model where sellers multihome than in the model where they single home (the bottleneck effect dominates the expansion effect). It follows that buyers are better off in the multihoming environment (as they benefit not only from larger seller participation but also from lower fees), while sellers who single home in both environments are worse off (as they pay higher fees for the same buyer participation); as for sellers who would multihome in the competitive bottleneck environment, we have shown above that they prefer the single-homing environment when $\tau_s < u$. So, in this case, buyers and sellers have diverging preferences regarding multihoming: sellers would prefer to be constrained to single home, while buyers would prefer that sellers were allowed to multihome.

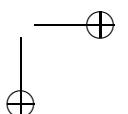
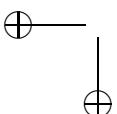
Second, if $\tau_s > u$, we can find parameter values for which all agents are better off in the competitive bottleneck environments. We already know that this is true for multihoming sellers (whatever the value of K). It is also true for single-homing sellers if $2\tau_s < K < 4\tau_s - 2u$ (as they pay lower fees in this case). Yet, in this region of parameters, buyers pay higher fees; the benefit of interacting with more sellers must then be sufficiently large to have $v_b^C > v_b^S$; this is so if $K > 2((2\pi + u)\tau_s - \pi u)/(\pi + u) \equiv K_0$, with $K_0 < 4\tau_s - 2u$ when $\tau_s > u$. In sum, if $\tau_s > u$ and $\max\{2\tau_s, K_0\} < K < 4\tau_s - 2u$, then both groups prefer the situation where sellers are allowed to multihome.

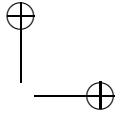
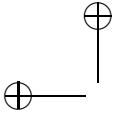
We can conclude that without further information, we cannot decide whether allowing multihoming on one side (with the other side single homing) leads to higher or lower net surpluses on either side. It is therefore a priori not possible to say whether the side that changes its behavior from single homing to multihoming (or the reverse) benefits or suffers from this change of behavior.

4.4 Media Markets

Media platforms are just one application of two-sided platforms where readers or viewers constitute one group and advertisers the other. A feature of most media platforms is that cross-group external effects go in opposite directions: while advertisers prefer a platform with more viewers, everything else given, the reverse holds true for viewers, as they dislike a platform that carries a lot of advertising. Thus, advertising is a nuisance. Many media platforms use a one-sided revenue model: they charge advertisers for posting ads and give away content bundled with advertising for free to viewers. This applies to free-to-air television, radio broadcasting and many media platforms on the Internet.

According to the baseline media model developed by Anderson and Coate (2005), two media platforms compete for viewers. While viewers are assumed to single home, advertisers





can post ads on multiple platforms. The key difference to the previous analysis is that media platforms are purely advertising financed. In particular, they are assumed to fix the advertising space or, equivalently, set the ad price per viewer.²³

Here, we review the basics of the competitive bottleneck model with n media platforms. Each media platform i provides program content to attract *viewers* and delivers these eyeballs (numbers of viewers) to *advertisers*.²⁴ Advertising revenue is the sole source of finance to platforms, and advertisers are assumed to be price takers. Platform i 's profit is thus $\Pi^i = P^i a^i$, $i = 1, \dots, n$, where P^i is the price per ad and a^i is the number of ads posted.

Content is attractive to viewers, but viewers consider the embodied ads to be a nuisance. Viewers' tastes over platform content is differentiated. Each viewer is assumed to single home, i.e., she makes a *discrete choice* over which platform to visit. Denote by $n_b^i(a^i, \mathbf{a}^{-i})$ the number of viewers (demand) for platform i as a function of its own ad level and the vector of ad levels, \mathbf{a}^{-i} , of its competitors.

On the advertiser side, all ads on a platform are seen by the viewers. Advertisers have different willingness to pay for reaching viewers (impressions). Assume that the advertiser's willingness to pay for advertising on each platform is a linear function of the number of viewers on the platform. In other words, there are constant returns to reaching prospective customers. This allows us to rank advertisers in terms of decreasing willingness to pay per eyeball, from large to small, giving a downward-sloping function $p(\cdot)$. When platform i opens a^i slots for advertising, the price per ad per viewer is $p(a^i)$, so that the price per ad is $P^i = p(a^i) n_b^i(a^i, \mathbf{a}^{-i})$. Hence, under these assumptions, we can write

$$\pi_i = a^i p(a^i) n_b^i(a^i, \mathbf{a}^{-i}) = R(a^i) n_b^i(a^i, \mathbf{a}^{-i}),$$

where $R(a^i)$ is the revenue per ad per viewer. The first-order condition (with ad levels as the strategic variables) is written as

$$\frac{R'(a^i)}{R(a^i)} = \frac{-\partial_1 n_b^i(a^i, \mathbf{a}^{-i})}{n_b^i(a^i, \mathbf{a}^{-i})}, \quad (11.2)$$

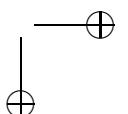
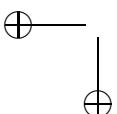
where ∂_1 is the partial derivative with respect to the first argument. This equation says that the elasticity of revenue per viewer should equal the viewer demand elasticity. This expression relates to the standard elasticity condition for oligopoly pricing.²⁵ The left-hand side, $R'(a)/R(a)$, is decreasing in a under the condition that $R(a)$ is log-concave. Hence, from (11.2), lower ad levels result whenever the equilibrium value of $-\partial_1 n_b^i(a^i, \mathbf{a}^{-i})/n_b^i(a^i, \mathbf{a}^{-i})$ increases in a change in market structure.

Consider the effects of platform entry at a symmetric equilibrium. For example, in the case of the Salop circle model, the right-hand side takes the value n/τ , where the transport

²³ This may approximate business practice. A standard practice in media markets is to report the CPM (cost per thousand impressions) based on past experience. Advertisers are then compensated if actual participation deviates from past participation.

²⁴ Our exposition follows Anderson et al. (2012). For a general treatment see Anderson and Peitz (2015).

²⁵ Indeed, consider the (Bertrand) oligopoly problem of $\max_{p_i} \pi_i = (p_i - c_i) n^i(p_i, \mathbf{p}_{-i})$ where now $n^i(p_i, \mathbf{p}_{-i})$ is the demand addressed to firm i and p_i is the price i sets for its product, while c_i is its marginal cost (and \mathbf{p}_{-i} is the vector of other firms' prices). Then the first-order condition can be written as $\frac{1}{(p_i - c_i)} = \frac{-N'_i(p_i, \mathbf{p}_{-i})}{N_i(p_i, \mathbf{p}_{-i})}$ which, in elasticity form, gives the inverse elasticity (Lerner) rule for pricing.



parameter τ measures how strongly platform content is differentiated. In the standard oligopoly context with product differentiation, this means simply that increasing the number of competitors leads to lower prices. In the present media economics context, this means that entry of a media platform leads to lower equilibrium ad levels. Competition for viewers plays out as competition in nuisance levels, and more competition leads to a lower nuisance level. For advertisers, the lower level of ads implies a *higher* ad price per viewer.

While the prediction of entry (and, relatedly, mergers) are unambiguous in this competitive bottleneck model, their empirical relevance may be questioned. Indeed, as Anderson et al. (2012) mention, at least two alternative mechanisms can overturn the result that entry reduces ad levels and that a merger increases ad levels. A countervailing effect arises when viewers spend some time on various media platforms and have limited attention for advertising they are exposed to. This introduces strategic interaction among platforms on the advertiser side. The attention of viewers becomes a common property resource that platforms tend to overexploit. A larger platform has stronger incentives to internalize the associated external effect on other ads and, therefore, opens fewer ad slots than a smaller platform. With symmetric platforms, the entry of a platform tends to lead to higher ad levels, as the negative effect of additional ads on existing advertisers is felt less strongly by each platform.²⁶

An alternative explanation that entry can lead to higher ad levels (and a merger to lower ad levels) rests purely on the effects of viewer multihoming on the exposure of viewers to ads. To the extent that a single impression is all that an advertiser cares about, advertising on multiple media carries the risk of wasting impressions because some viewers will be treated twice. This has implications for the advertising strategy of media platforms (as well as the decision on content of platforms).

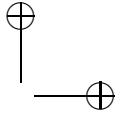
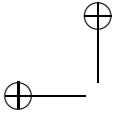
Ambrus, Calvano, and Reisinger (2016) consider a setting with viewers who can multihome and homogeneous advertisers who can post multiple ads.²⁷ Start with a monopoly setting and suppose that a second platform enters. There are two effects due to entry. First, there is a duplication effect: as each multihoming consumer can now get informed about an advertiser's product on both platforms, the single ad is worth less. Due to the duplication effect, the advertising intensity tends to fall in duopoly. However, there is a countervailing effect, which can be called the business-sharing effect. In monopoly, all consumers are exclusive consumers. By contrast, in duopoly, some of the lost business due to increased advertising levels comes from consumers active on both platforms. Here, the duopolist shares business with its rival. Losing these consumers is less detrimental than losing exclusive consumers. Due to this business-sharing effect, a duopolist tends to have a greater incentive to increase the advertising level than a monopolist. The business-sharing effect possibly dominates the duplication effect, in which case advertising levels in duopoly are larger than in monopoly.

4.5 Further Issues

The focus of our presentation in this section has been to determine features of the equilibrium allocation of markets in which one or multiple platforms enable the interaction between two

²⁶ For a formal analysis of media markets with limited viewer attention for ads, see Anderson and Peitz (2016).

²⁷ For a related contribution, see Anderson, Foros, and Kind (2015). For an overview of the effects of viewer multihoming in media markets, see Peitz and Reisinger (2015).



groups of users, restricting the analysis to stylized and simple settings. Here, we look at, in some sense, richer settings: (i) richer price instruments by the platform, (ii) within-group external effects and, in particular, competition among sellers, and (iii) investment incentives by sellers and platforms.

4.5.1 Price instruments

In most of the analysis we postulated that platforms charge listing or access fees to the platform. However, pricing strategies may be more involved. In particular, platforms may choose two-part tariffs or offer some insurance against unexpected drops in participation on the other side.

Two-part tariffs Reisinger (2014) allows platforms to set two-part tariffs, adding a usage (or per-transaction) fee to the membership (or subscription) fee that we considered so far. It is not unusual that platforms charge two-part tariffs to at least one of the sides. For instance, in software platforms, developers are charged a fixed fee for getting access to source code of the system and, in addition, pay royalties for the applications they sell to users. What are the implications of this form of price discrimination on the profits of competing platforms and on the welfare of the two sides?

The model is modified as follows. The seller and buyer surpluses gross of any opportunity cost of visiting a platform become

$$v_s^i = n_b^i (\pi - t_s^i) - m_s^i \quad \text{and} \quad v_b^i = n_s^i (u - t_b^i) - m_b^i,$$

where t_k^i is the transaction (usage) fee that platform i charges on side k . As for platform i 's profit, it is now expressed as

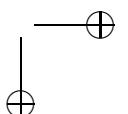
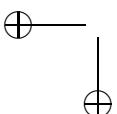
$$\Pi^i = (m_s^i - c_s) n_s^i + (m_b^i - c_b) n_b^i + (t_s^i + t_b^i - \gamma) n_s^i n_b^i,$$

where γ is the constant per-transaction cost incurred by the platform. The game is the same as before, except that platforms now have four strategic variables to choose.

A general result that emerges is that a continuum of equilibria exists when platforms compete using two-part tariffs.²⁸ The reason behind this multiplicity is that platforms only care about the *total* price that agents pay but not about how it splits between the membership and the transaction fees (i.e., different combinations of the two fees yield the same profit); as a result, platforms have a continuum of best responses to their rival's tariff. For instance, in the two-sided single-homing game, the continuum of symmetric equilibria is characterized by platforms charging $T_k = m_k + t_k n_k$ ($k = s, b$), where $m_s = c_s + \tau_s - u + \frac{1}{2}(t_b - t_s)$, $m_b = c_b + \tau_b - \pi + \frac{1}{2}(t_s - t_b)$ and $0 \leq m_s \leq 2\pi, 0 \leq m_b \leq 2u$. The platforms' equilibrium profits are given by $\Pi = \frac{1}{2}(\tau_s + \tau_b - \pi - u) + \frac{1}{4}(t_s + t_b)$.

As the previous expressions clearly show, these equilibria lead to different profits for platforms and different surpluses for agents. An unwelcome consequence is that the models (with two-sided single homing or with competitive bottlenecks) are deprived of any predictive power.

²⁸ This was already pointed out by Armstrong (2006).



To solve this problem, Reisinger (2014) allows for heterogeneous trading behavior among agents on both sides. In particular, there exist two types of agents on each side: the ‘normal’ agents interact with all agents on the other side (as was assumed so far), while the ‘small’ agents only interact with a *fraction* of the agents on the other side (or interact with each of them only with some probability); it is further assumed that the small agents are a minority and that platforms are unable to price discriminate across types. In the competitive bottlenecks model where sellers can multihome while buyers cannot, Reisinger shows that this formulation leads to a unique equilibrium in the price game, even when the masses of small agents become infinitely small (i.e., when heterogeneity disappears). Moreover, this equilibrium has many reasonable properties; in particular, it is still the case that platforms price aggressively to the side that exerts the larger cross-group external effect; the difference under two-part tariffs is that the lower payment is only reflected in the transaction fee but not in the membership fee.²⁹

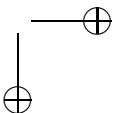
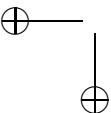
The intuition for the unique result is the following. The two types of agents react differently to a particular combination of membership and transaction fees. To keep the utility of a small agent (who trades less often) constant, platforms have to balance an increase in the transaction fee with a *smaller* reduction of the membership fee than they would do for a normal seller. It follows that the effect on profit of a marginal change in platform i ’s transaction fee is no longer a constant multiple of the effect of a marginal change in i ’s membership fee. This multiple varies continuously as the fees change because the ratio of the two types that join platform i also varies continuously. As a result, each platform has a unique optimal combination of the fees as a reaction to the price quadruple of its rival.

Insulating tariffs An alternative approach to platform pricing is to consider that platforms set so-called ‘insulating tariffs’, i.e., tariffs such that after a deviation on one side, the pricing on the other side is adjusted in order to insulate the demand effect on one side and leave demand on the other side unchanged. This concept has been proposed and developed by Weyl (2010) in a monopoly setting, and Weyl and White (2016) in a duopoly setting.

Suppose that platforms adapt their pricing to insure agents against any utility loss from low participation on the other side of the market. Let us apply this alternative strategy space to our previous model of two-sided single homing. Instead of setting the membership fees m_s^i and m_b^i , platforms now set the net surpluses v_s^i and v_b^i . This change of strategic variables removes the feedback effect stemming from cross-group external effects. We should therefore expect a weaker impact of cross-group effects on equilibrium prices. We check now that this conjecture is correct.

Recalling that $v_s^i = r_s + \pi n_b^i - m_s^i$ and $v_b^i = r_b + un_s^i - m_b^i$, we can write membership fees as a function of participation on the other side and utilities v_s^i and v_b^i : $m_s^i = r_s + \pi n_b^i - v_s^i$ and $m_b^i = r_b + un_s^i - v_b^i$. If a platform sets an insulating tariff (or offers a guarantee in utils) to all agents, it fixes v_s^i and v_b^i , which means that agents are insured against any changes in platform participation on the other side since the membership fee will be adjusted accordingly. We can then write platform i ’s profit as $\Pi^i = (r_s + \pi n_b^i - v_s^i - c_s)n_s^i + (r_b + un_s^i - v_b^i - c_b)n_b^i$. Recalling

²⁹ For a discussion of the link between heterogeneity and price instruments, see Rochet and Tirole (2006).



that the participation levels n_s^i and n_b^i are obtained in the Hotelling fashion and can thus be expressed as a function of the guarantees, we have

$$\begin{aligned}\Pi^i &= \left[r_s + \pi \left(\frac{1}{2} + \frac{v_b^i - v_b^j}{2\tau_b} \right) - v_s^i - c_s \right] \left(\frac{1}{2} + \frac{v_s^i - v_s^j}{2\tau_s} \right) \\ &\quad + \left[r_b + u \left(\frac{1}{2} + \frac{v_s^i - v_s^j}{2\tau_s} \right) - v_b^i - c_b \right] \left(\frac{1}{2} + \frac{v_b^i - v_b^j}{2\tau_b} \right).\end{aligned}$$

We can now characterize the Nash equilibrium of the game in which platforms simultaneously choose guarantees. This allows us to determine the membership fees that platforms charge at equilibrium:³⁰

$$m_s^* = c_s + \tau_s - u/2, \text{ and } m_b^* = c_b + \tau_b - \pi/2.$$

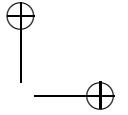
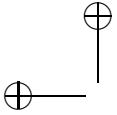
Recall that in the previous model in which platforms choose membership fees, equilibrium membership fees are equal to $c_s + \tau_s - u$ and $c_b + \tau_b - \pi$, which are lower than the ones obtained here. This confirms our initial intuition that the impact of cross-side external effects on competition is weaker when platforms offer guarantees, as there are no feedback effects on platform participation from one side to the other. The lesson to emerge from the comparison of these two models is that platforms that can compensate for utility losses due to lower than promised participation on the other side will compete less fiercely than platforms that only choose prices.

4.5.2 Within-group external effects

The analyses presented thus far considered exclusively cross-group external effects. Such a focus seems natural as cross-group effects directly stem from the desire of the two groups of agents to interact and, thereby, give their raison d'être to two-sided platforms. However, there are a number of situations where platforms also have to factor in the existence of within-group external effects when choosing their strategies. As explained in the introduction, these effects describe the fact that the attractiveness of a platform for the members of one group depends on the participation of the members of the very same group. Within-group effects are negative when the members of one group compete with one another to interact with the other group (e.g., Uber drivers face a given set of passengers at any location and any point in time) or because of some form of congestion (e.g., talking about Airbnb visitors, Slee, 2016 reports that 'as their numbers grow, they erode the very atmosphere in which they bask and threaten the livability of the city for residents'). In contrast, there also exist sources of positive within-group effects; for instance, Belleflamme, Omrani, and Peitz (2015) explain that a larger 'crowd' of funders on a crowdfunding platform increases the probability that any project will be realized, which benefits all funders. In what follows, we focus on markets with competing platforms.³¹ We first generalize the two-sided single homing model of Armstrong (2006) to

³⁰ See Belleflamme and Peitz (2015, pp. 671–672) for the details. Note that in contrast with the model where platforms set membership fees, an equilibrium with two active platforms exists here without having to impose that platforms be sufficiently differentiated.

³¹ For an analysis with within-group external effects on monopoly platforms, see Nocke, Peitz, and Stahl (2007). They revisit the question of socially excessive or insufficient entry of monopolistically competitive sellers when the



allow for any type of cross- and within-group external effects. We then examine the influence of within-group effects on the coexistence of platforms.

A two-sided single-homing model encompassing within-group effects Following Belleflamme and Toulemonde (2016b), we define the seller and buyer net surplus of visiting platform i (gross of any opportunity cost) respectively as $v_s^i = r_s + \pi(n_b^i, n_s^i) - m_s^i$ and $v_b^i = r_b + u(n_b^i, n_s^i) - m_b^i$. The difference with the previous model is that we use now the general functions $\pi(n_b^i, n_s^i)$ and $u(n_b^i, n_s^i)$ to represent the net gains from trade for any seller and any buyer on platform i . They both potentially depend on the number of buyers and on the number of sellers who are present on the platform, meaning that any form of cross-group and of within-group external effects are permitted. Both functions are supposed to be twice continuously differentiable in their two arguments. We proceed as before by identifying the indifferent seller and buyer in the standard Hotelling fashion, which allows us to express the numbers of sellers and buyers at platform i as:

$$\begin{cases} n_b^i = \frac{1}{2} + \frac{1}{2\tau_b} \Delta^u(n_b^i, n_s^i) - \frac{1}{2\tau_b} (m_b^i - m_b^j), \\ n_s^i = \frac{1}{2} + \frac{1}{2\tau_s} \Delta^\pi(n_b^i, n_s^i) - \frac{1}{2\tau_s} (m_s^i - m_s^j), \end{cases} \quad (11.3)$$

where

$$\begin{aligned} \Delta^u(n_b^i, n_s^i) &\equiv u(n_b^i, n_s^i) - u(1 - n_b^i, 1 - n_s^i), \\ \Delta^\pi(n_b^i, n_s^i) &\equiv \pi(n_b^i, n_s^i) - \pi(1 - n_b^i, 1 - n_s^i). \end{aligned}$$

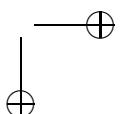
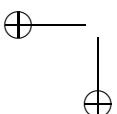
We also introduce the following notation:

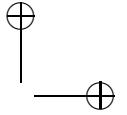
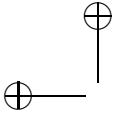
$$\begin{aligned} \Delta_b^u &\equiv \frac{\partial \Delta^u(n_b^i, n_s^i)}{\partial n_b^i}, \Delta_s^u \equiv \frac{\partial \Delta^u(n_b^i, n_s^i)}{\partial n_s^i}, \\ \Delta_b^\pi &\equiv \frac{\partial \Delta^\pi(n_b^i, n_s^i)}{\partial n_b^i}, \Delta_s^\pi \equiv \frac{\partial \Delta^\pi(n_b^i, n_s^i)}{\partial n_s^i}. \end{aligned}$$

In words, the function $\Delta^u(n_b^i, n_s^i)$ measures the differential in buyers' net gains from trade between platforms i and j when there are n_b^i buyers and n_s^i sellers on platform i . The derivatives Δ_b^u and Δ_s^u measures the sensitivity of this differential to a change in the mass of, respectively, buyers or sellers on platform i ; the function $\Delta^\pi(n_b^i, n_s^i)$ and derivatives Δ_b^π and Δ_s^π are defined accordingly for sellers.

The system of equations (11.3) implicitly determines the demand functions for platform i , $n_b^i(m_b^i, m_s^i, m_b^j, m_s^j)$ and $n_s^i(m_b^i, m_s^i, m_b^j, m_s^j)$, which depend on the combination of the four

monopoly platform can charge only sellers for their listing service, and compare their findings to what happens with alternative governance structures of the platform.





fees.³² Using implicit differentiation and taking advantage of the fact that $n_s^1 = n_s^2 = n_b^1 = n_b^2 = \frac{1}{2}$ at the symmetric equilibrium, it is then possible to show that the platforms set the following membership fees at the symmetric equilibrium of the game:

$$m_s^* = c_s + \tau_s - \frac{1}{2} (\Delta_s^u + \Delta_s^\pi),$$

$$m_b^* = c_b + \tau_b - \frac{1}{2} (\Delta_b^u + \Delta_b^\pi),$$

where Δ_b^u , Δ_b^π , Δ_s^u , and Δ_s^π are evaluated at $(n_b^i, n_s^i) = (\frac{1}{2}, \frac{1}{2})$. We observe that the equilibrium membership fees depend on the nature of the within- and cross-group external effects. In the complete absence of external effects within and across groups, fees would be as in the Hotelling model. The presence of positive (resp. negative) external effects from, say, sellers to buyers, leads platforms to lower (resp. raise) the membership fee for sellers below (resp. above) the level that would prevail absent any external effect. This is the standard result of Armstrong (2006). We add here a new result related to the presence of external effects *within* groups. Positive (resp. negative) external effects within groups leads platforms to lower (resp. raise) the membership fee for the group below (resp. above) the level that would prevail absent any external effect.³³

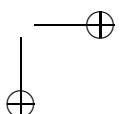
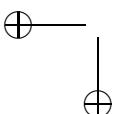
Coexistence of platforms An interesting issue is the impact that within-side external effects may have on the coexistence of competing two-sided platforms. We have seen above that positive cross-group effects generate positive feedback loops that may lead to situations where only one platform survives at equilibrium ('winner-takes-all') unless competing platforms are sufficiently differentiated. We may conjecture, however, that negative within-side effects may contribute to break the feedback loop and, thereby, facilitate the coexistence of competing platforms, even in the absence of differentiation.

Karle, Peitz, and Reisinger (2016) address this issue by examining how the degree of competition among sellers affects the possibility for non-differentiated platforms offering listing services to coexist at equilibrium.³⁴ Recall that Belleflamme and Toulemonde (2016b) defined the seller and buyer net surplus of visiting platform i respectively as $v_s^i = r_s + \pi(n_b^i, n_s^i) - m_s^i$ and $v_b^i = r_b + u(n_b^i, n_s^i) - m_b^i$. A special case is that the profit per buyer depends on the number competing sellers on the platform and platforms only charge sellers, $v_s^i = r_s + n_b^i \pi(n_s^i) - m_s^i$ and $v_b^i = r_b + n_s^i u$. Using these surplus functions, Karle et al. (2016) analyze the two-sided single-homing model in which buyers observe product offerings on a platform only after having visited the platform. If all sellers co-locate on the same platform then, in equilibrium, all buyers will be active on this platform. Thus there is agglomeration in

³² It is assumed that the functions u and π are such that the system (11.3) leads to a unique solution $(n_b^i, n_b^j) \in [0, 1]^2$, which is well-behaved in the sense that both n_b^i and n_b^j are decreasing functions of $(m_b^i - m_b^j)$ and $(m_s^i - m_s^j)$.

³³ To be sure, we recover the previous results by setting $\pi(n_b^i, n_s^i) = \pi n_b^i$ and $u(n_b^i, n_s^i) = u n_s^i$ (i.e., cross-group effects are positive and linear and within-group effects are nil). We have then $\Delta^\pi = \pi(2n_b^i - 1)$, $\Delta^u = u(2n_s^i - 1)$, $\Delta_b^\pi = 2\pi$, $\Delta_s^u = 2u$, and $\Delta_s^\pi = \Delta_b^u = 0$. It follows that $m_s^* = c_s + \tau_s - u$ and $m_b^* = c_b + \tau_b - \pi$.

³⁴ Ellison and Fudenberg (2003) and Ellison, Fudenberg, and Möbius (2004) also show that negative within-side effects may contribute to the coexistence of two-sided platforms. Hagiu (2009) introduces competition among sellers on a two-sided platform; competition stems from consumers' variable preference for variety over sellers' products, which turns out to be a key factor determining the optimal pricing structure either of a monopoly platform or of competing platforms.



equilibrium and network effects are fully exploited. In this equilibrium all profits are competed away and both platforms make zero profits.

However, imperfect competition between sellers may lead to a different equilibrium in which both platforms have a positive number of users and make positive profits in equilibrium. Suppose there are two sellers that have to decide whether to join platform 1, join platform 2 or not to participate at all. If they both join the same platform, they obtain duopoly profit π^d per buyer, which is less than monopoly profit π^m per buyer they would obtain if they were the only seller on the platform. If π^m is sufficiently large, there is an equilibrium in which sellers list on different platforms. Buyers are indifferent between the two platforms. Profits are not competed away; platforms can extract the full seller surplus in equilibrium.

Belleflamme and Toulemonde (2009) propose another way to address this issue. They examine the extent to which negative within-group effects among sellers may help a new platform operator lure buyers and sellers away from an existing platform. In their model, only the new platform can set membership fees; there is no price competition per se, but this is not a monopoly setting either as the existing platform provides buyers and sellers with an outside option. As in Caillaud and Jullien (2003), the new platform faces a ‘chicken-and-egg’ problem, which it tries to solve by using a divide-and-conquer pricing strategy; that is, the platform must subsidize the participation of one side (divide) and hope to recoup the loss through the membership fee it sets on the other side (conquer). The question is whether the platform can make any profit with such a strategy. The answer is yes when the interaction among buyers and sellers only generates (positive) cross-group external effects. However, the presence of negative within-group effects among sellers (e.g., because they offer substitutable products) blurs the picture. Competition among sellers turns out to be a mixed blessing for the new platform. The upside is that the sellers’ willingness to pay to join the new platform increases if only a few of them make the move; as a consequence, sellers are less sensitive to buyers’ participation to the new platform, which alleviates the ‘chicken-and-egg’ problem. Yet, the downside is that it will be more costly for the new platform to attract buyers if only a small subset of the sellers join. The balance between the two effects depends on the relative strength of the within-group effects (with respect to the cross-group effects). There may be situations where entry is not profitable.

4.5.3 Investment issues

So far, the competition among platforms was studied in fixed environments. All the models we considered implicitly assumed that neither platforms nor their users are able to modify the conditions under which the interaction among the two groups is taking place. We now relax this assumption in two different ways: first, we let sellers make ex ante investments that affect the gains from trade when they interact with buyers; second, we let platforms invest in reducing their costs of registering agents.

Seller investment incentives on a platform Belleflamme and Peitz (2010) analyze how seller investment incentives are affected by the presence of competing for-profit platforms. Platform competition is modeled as in Armstrong (2006) and Armstrong and Wright (2007); that is, the final stages of the game are the ones we analyzed above with either both sides single homing, or one or the other side being allowed to multihome. The novelty is to add an initial stage where sellers have the possibility to make long-term investments, which may take the form of cost reduction, quality improvement or marketing measures that facilitate price discrimination

or expand demand. These investments affect the surpluses, π and u , that sellers and buyers obtain when they trade on the platforms. Typically, investments in cost reduction or quality improvement increase both π and u , while investments in better price discrimination increase π while decreasing u .

To assess the impact that for-profit platforms have on investment incentives, two trading environments are contrasted. The first is the one we are considering in this section and is called ‘intermediated trade’: trade takes place through for-profit intermediaries, which set membership fees on both sides of the market. The other environment is called ‘non-intermediated trade’, as trade is assumed to take place via open trading platforms, which can be accessed without charge.

One could think that seller investment incentives would be weaker in the intermediated trade environment, simply because the for-profit intermediaries capture part of the rents that are available in the market. However, this reasoning is short-sighted because investments also affect the membership fees that platforms set at equilibrium. Why? Because seller investments modify the intensity of the cross-group external effects and, thereby, the competition between the platforms. In particular, investments that increase buyer surplus lead platforms to lower their fee on the seller side. As a consequence, sellers internalize changes in buyer surplus if products are traded on for-profit platforms, whereas they do not in the context of open platforms. It follows that investment incentives can be stronger with competing for-profit platforms than with open platforms. The exact relationship between investment incentives and for-profit intermediation depends on which side of the market single homes and on the nature of the investment effort. In general, it can be said that as the intensity of competition for sellers increases, proprietary platforms are more likely to provide better seller investment incentives than open platforms. Indeed, this happens when the nature of platform competition moves from multihoming sellers and single-homing buyers to single-homing sellers and buyers, and then to single-homing sellers and multihoming buyers.

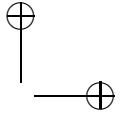
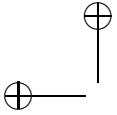
Platform investment incentives To study the investment incentives of competing platforms, we need first to solve a game of price competition among asymmetric platforms. Belleflamme and Toulemonde (2016a) do so for the two-sided single-homing model of Armstrong (2006). We present here a simplified version of their work by letting only marginal costs differ across platforms (all the other parameters, in particular, the cross-group external effects, remain common to the two platforms). Let c_s^i and c_b^i denote the marginal cost for platform i of registering, respectively, an extra seller and an extra buyer; let also $\gamma_s \equiv c_s^i - c_s^j$ and $\gamma_b \equiv c_b^i - c_b^j$ ($i, j = 1, 2; i \neq j$). Repeating all the steps described in Section 4.3.1, they find the following equilibrium membership fees for the two platforms ($i, j = 1, 2; i \neq j$):

$$\begin{aligned} m_s^{i*} &= c_s^i + \tau_s - u - \frac{1}{3}\gamma_s - \frac{1}{3D}(\pi - u)[(2\pi + u)\gamma_s + 3\tau_s\gamma_b], \\ m_b^{i*} &= c_b^i + \tau_b - \pi - \frac{1}{3}\gamma_b - \frac{1}{3D}(u - \pi)[(2u + \pi)\gamma_b + 3\tau_s\gamma_s], \end{aligned}$$

where $D \equiv 9\tau_s\tau_b - (2\pi + u)(\pi + 2u)$ is positive.³⁵

The equilibrium membership fees can be decomposed as the sum of four components: (i) the first two terms ($c_k^i + \tau_k$) are the classic Hotelling formula (marginal cost + transportation

³⁵ This follows from the second-order condition $4\tau_s\tau_b > (\pi + u)^2$.



cost); (ii) the third term was identified by Armstrong (2006) as the price adjustment due to cross-group external effects (the fee is decreased by the externality exerted on the other side); (iii) the fourth term is the effect of cost differences across platforms; (iv) the last term results from the interplay between cost differences and cross-group external effects. If platforms are symmetric ($\gamma_k = 0$), we find the same formulas as in Section 4.3.1. In the particular case where cross-side external effects are the same on the two sides ($\pi = u$, meaning here that the gains from trade are equally split among buyer and seller), all terms but the last remain. The latter result is reminiscent of what we already observed in the setting with a monopoly platform.

The equilibrium profit of platform i is computed as

$$\begin{aligned}\Pi^{i*} = & \frac{1}{2} (\tau_s + \tau_b - \pi - u) + \frac{1}{2D} \left(\tau_b \gamma_s^2 + \tau_s \gamma_b^2 \right) + \frac{1}{2D} (\pi + u) \gamma_s \gamma_b \\ & - \frac{\gamma_s}{2D} (6\tau_s \tau_b + \tau_b (\pi - u) - (\pi + u) (2\pi + u)) \\ & - \frac{\gamma_b}{2D} (6\tau_s \tau_b - \tau_s (\pi - u) - (\pi + u) (\pi + 2u)),\end{aligned}\quad (11.4)$$

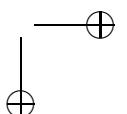
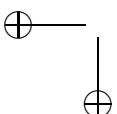
with Π^{j*} being obtained by replacing γ_s by $-\gamma_s$, and γ_b by $-\gamma_b$.

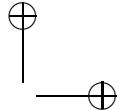
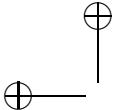
Belleflamme and Toulemonde (2016a) then use this equilibrium profit function to estimate platform i 's incentives to invest in cost reduction. Their main contribution is to show that cross-group external effects affect incentives to invest in cost reduction through the *strategic effect* of this investment. The strategic effect is the effect on one platform's profit that operates through the modification of the other platform's equilibrium fees. Absent cross-group external effects, we expect the strategic effect of a lower cost to be negative if firms compete in prices over substitutable services: a lower cost for firm A leads this firm to decrease its price, which leads the rival firm to decrease its price as well (because of strategic complementarity); this, in turn, reduces firm A 's profit, which contributes to attenuating the direct positive impact of cost reduction on profits.

The presence of cross-group external effects challenges the previous results in two major ways. First, cross-group external effects may decrease the strategic effect and they may do so to such an extent that the strategic effect outweighs the positive direct effect; it follows that the net effect of lower costs on profit becomes negative. In that case, platforms would be better off if they could increase, rather than decrease, their costs. Second, in complete contrast with the previous case, external effects may increase the strategic effect, even up to a point where it becomes positive; in the latter case, platforms would have a twofold incentive to invest in cost reduction as it would benefit them first directly and next, indirectly, through the upward adjustment of the rival platform's equilibrium prices. It is shown that for either of these extreme cases to arise, cross-group external effects must be large relatively to the intensity of competition on the two sides.

5 CONCLUSION

In intermediate microeconomics, students learn that market power, asymmetric information, and externalities are sources of market failure. Classic oligopoly theory has focused on market power; modern industrial organization has, in addition, incorporated asymmetric





316 *Handbook of game theory and industrial organization: applications*

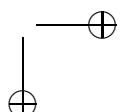
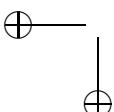
information into the analysis of markets. While markets with network effects have also been investigated by industrial organization economists at least since the beginning 1980s, more recent efforts investigate decision making by intermediaries in markets characterized by imperfect competition and external effects. A particular focus is on market environments in which a platform caters to multiple audiences, which are distinct, but connected.

A number of theoretical insights have been derived, some of which are reviewed in this chapter. This chapter has been silent regarding the empirical literature on network effects and platforms. An early guide to empirical work on two-sided platforms is Rysman (2009). Some considerable work has been done in the context of media markets; we refer the reader to the overviews by Chandra and Kaiser (2015), Crawford (2015), and Sweeting (2015).

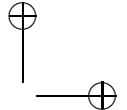
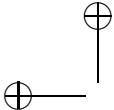
This chapter has ignored recent work on competition policy issues regarding two-sided platforms. A case in point is the analysis of price-parity clauses imposed by platforms (see, in particular, Edelman and Wright, 2015). Also, there is a well-developed literature on payment systems in which the two-sidedness of the payment system is a critical feature. Finally, the two-sidedness of the business model features prominently in the economic analysis of net neutrality (see Greenstein, Peitz, and Valletti, 2016). We expect the analysis of two-sided platforms to be one of the most active areas of research in industrial organization theory in the years to come.

REFERENCES

- Ambrus, A., E. Calvano, and M. Reisinger (2016). Either or Both Competition: A ‘Two-Sided’ Theory of Advertising with Overlapping Viewerships. *American Economic Journal: Microeconomics* 8, 189–222.
- Anderson, S. and S. Coate (2005). Market Provision of Broadcasting: A Welfare Analysis. *Review of Economic Studies* 72, 947–972.
- Anderson, S. and M. Peitz (2015). Media See-saws: Winner and Losers on Media Platforms, *University of Mannheim Working Paper* 15–16.
- Anderson, S. and M. Peitz (2016). Advertising Congestion in Media Markets. Unpublished manuscript.
- Anderson, S., O. Foros and H.-J. Kind (2015). Competition for Advertisers and for Viewers in Media Markets. *CEPR Discussion Paper* 10608.
- Anderson, S., O. Foros, H.-J. Kind, and M. Peitz (2012). Media Market Concentration, Advertising Levels, and Ad Prices. *International Journal of Industrial Organization* 30, 321–325.
- Armstrong, M. (1998). Network Interconnection in Telecommunications. *Economic Journal* 108, 545–564.
- Armstrong, M. (2006). Competition in Two-sided Markets. *Rand Journal of Economics* 37, 668–691.
- Armstrong, M. and J. Wright (2007). Two-sided Markets, Competitive Bottlenecks and Exclusive Contracts. *Economic Theory* 32, 353–380.
- Belleflamme, P. and M. Peitz (2010). Platform Competition and Seller Investment Incentives. *European Economic Review* 54, 1059–1076.
- Belleflamme, P. and M. Peitz (2015). *Industrial Organization: Markets and Strategies*. 2nd edition. Cambridge, UK: Cambridge University Press.
- Belleflamme, P. and M. Peitz (2016). Platform Competition: Who Benefits from Multihoming? Unpublished manuscript, University of Mannheim.
- Belleflamme, P. and E. Toulemonde (2009). Negative Intra-group Externalities in Two-sided Markets. *International Economic Review* 50, 245–272.
- Belleflamme, P. and E. Toulemonde (2016a). Tax Incidence on Competing Two-sided Platforms: Lucky Break or Double Jeopardy. *CORE Discussion Paper* 2016/12.
- Belleflamme, P. and E. Toulemonde (2016b). Who Benefits from Increased Competition among Sellers on B2C Platforms? *Research Economics* 70, 741–751.
- Belleflamme, P., N. Omrani, and M. Peitz (2015). The Economics of Crowdfunding Platforms. *Information Economics and Policy* 33, 11–28.
- Cabral, L. (2011). Dynamic Price Competition with Network Effects. *Review of Economic Studies* 78, 83–111.
- Caillaud, B. and B. Jullien (2001). Competing Cybermediaries. *European Economic Review (Papers and Proceedings)* 45, 797–808.



- Caillaud, B. and B. Jullien (2003). Chicken and Egg: Competition Among Intermediation Service Providers. *Rand Journal of Economics* 34, 521–552.
- Chandra, A. and U. Kaiser (2015). Newspapers and Magazines, in: S.P. Anderson, D. Stromberg, and J. Waldfogel (eds), *Handbook of Media Economics, Vol. 1A*, Amsterdam: Elsevier, pp. 397–444.
- Choi, J.P. (1994). Network Externality, Compatibility Choice, and Planned Obsolescence. *Journal of Industrial Economics* 42, 167–182.
- Crawford, G. (2015). The Economics of Television and Online Video Markets, in S.P. Anderson, D. Stromberg, and J. Waldfogel (eds), *Handbook of Media Economics, Vol. 1A*, Amsterdam: Elsevier, pp. 267–340.
- Crémier, J., P. Rey, and J. Tirole (2000). Connectivity in the Commercial Internet. *Journal of Industrial Economics* 48, 433–472.
- De Bijl, P. and M. Peitz (2002). *Regulation and Entry into Telecommunications Markets*. Cambridge, UK: Cambridge University Press.
- Doganoglu, T. and J. Wright (2006). Multihoming and Compatibility, *International Journal of Industrial Organization* 24, 45–67.
- Edelman, B. and J. Wright (2015). Price Coherence and Excessive Intermediation, *Quarterly Journal of Economics* 130, 1283–1328.
- Ellison, G. and D. Fudenberg (2003). Knife-edge or Plateau: When Do Market Models Tip? *Quarterly Journal of Economics* 118, 1249–1278.
- Ellison, G., D. Fudenberg, and M. Möbius (2004). Competing Auctions. *Journal of the European Economic Association* 2, 30–66.
- European Commission (2015). *Public Consultation on the Regulatory Environment for Platforms, Online Intermediaries, Data and Cloud Computing and the Collaborative Economy*.
- Fujita, M. and J.F. Thisse (2013). *Economics of Agglomeration. Cities, Industrial Location, and Globalization*. 2nd edition. Cambridge, UK: Cambridge University Press.
- Greenstein, S., M. Peitz, and T. Valletti (2016). Net Neutrality: A Fast Lane to Understanding the Trade-offs. *Journal of Economic Perspectives* 30, 127–149.
- Grilo, I., O. Shy, and J.F. Thisse (2001). Price Competition When Consumer Behavior Is Characterized by Conformity or Vanity, *Journal of Public Economics* 80, 385–408.
- Hagiu, A. (2009). Two-sided Platforms: Product Variety and Pricing Structures. *Journal of Economics and Management Strategy* 18, 1011–1043.
- Hoernig, S., R. Inderst, and T. Valletti (2014). Calling Circles: Network Competition with Nonuniform Calling Patterns. *Rand Journal of Economics* 45, 155–175.
- House of Lords (2016). *Online Platforms and the Digital Single Market*. Report published April 20, 2016.
- Karle, H., M. Peitz, and M. Reisinger (2016). Segmentation versus Agglomeration: Competition Between Platforms with Competitive Sellers. Mimeo.
- Katz, M. and C. Shapiro (1985). Network Externalities, Competition and Compatibility. *American Economic Review* 75: 424–440.
- Laffont, J.-J., P. Rey, and J. Tirole (1998a). Network Competition: I. Overview and Nondiscriminatory Pricing. *Rand Journal of Economics* 29, 1–37.
- Laffont, J.-J., P. Rey, and J. Tirole (1998b). Network Competition: II. Price Discrimination. *Rand Journal of Economics* 29, 38–56.
- Marshall, A. (1890 [1920]). *Principles of Economics*. 8th edition. London: Macmillan.
- Monopolkommission (2015). *Competition Policy: The Challenge of Digital Markets*. Special Report by the Monopolies Commission, June 1, 2015.
- Nocke, V., M. Peitz, and K. Stahl (2007). Platform Ownership. *Journal of the European Economic Association* 5, 1130–1160.
- Peitz, M. and M. Reisinger (2015). Media Economics of the Internet, in: S.P. Anderson, D. Stromberg, and J. Waldfogel (eds), *Handbook of Media Economics, Vol. 1A*, Amsterdam: Elsevier, pp. 445–530.
- Reisinger, M. (2014). Two-part Tariff Competition between Two-sided Platforms. *European Economic Review* 68, 168–180.
- Rochet, J.-C. and J. Tirole (2003). Platform Competition in Two-sided Markets, *Journal of the European Economic Association* 1, 990–1024.
- Rochet, J.-C. and J. Tirole (2006). Two-sided Markets: A Progress Report. *Rand Journal of Economics* 37, 645–667.
- Rysman, M. (2009). The Economics of Two-sided Markets. *Journal of Economic Perspectives* 23, 125–143.
- Slee, T. (2016). Airbnb Is Facing an Existential Expansion Problem. *Harvard Business Review*. Available at <https://hbr.org/2016/07/airbnb-is-facing-an-existential-expansion-problem>.
- Sweeting, A. (2015). Radio, in S.P. Anderson, D. Stromberg, and J. Waldfogel (eds), *Handbook of Media Economics, Vol. 1A*, Amsterdam: Elsevier, pp. 341–396.
- Weyl, E.G. (2010). A Price Theory of Multi-sided Platforms. *American Economic Review* 100, 1642–1672.
- Weyl, E.G. and A. White (2016). Insulated Platform Competition. Mimeo. Available at <http://ssrn.com/abstract=1694317>.



12. Auctions

*Ángel Hernando-Veciana**

1 INTRODUCTION

Auction theory is a field of economics that has been developed during the last 50 years starting from Vickrey's (1961) seminal paper. Its evolution has been one of the most successful applications of game theory with remarkable episodes like the spectrum auctions in the 1990s; see Cramton (1995) and Binmore and Klemperer (2002).

Nowadays, auction theory is a well-established branch of economics for which there exist influential handbooks covering the basic tools and applications, e.g. Krishna (2002), Milgrom (2004), Klemperer (2004) and Menezes and Monteiro (2008). In this survey, I am going to provide a revision of advances in the field that have taken place around the last decade. To do so, I will structure this survey as follows. It starts with an introductory section in which I summarize the basic tools that I refer to in the subsequent sections. Next, I emphasize the advances in three innovative areas: position auctions, Internet auctions and combinatorial auctions. I finish with a section summarizing some remarkable contributions to auction theory throughout the decade organized by topics.

2 ELEMENTARY AUCTION THEORY

In this section, I revise the central contributions of auction theory with special emphasis on the revenue equivalence theorem and the Vickrey-Clarke-Groves auctions. These are central results that I will use in my discussions in the next sections.

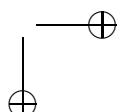
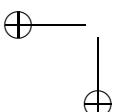
The usual approach in auction theory is to study a game of asymmetric information whose rules reflect a particular real-life auction. Formally, the setting is described by a set of n bidders, $i \in I \equiv \{1, \dots, n\}$, each with a payoff function $u_i(o, \theta_i, \theta_{-i})$ where o denotes the outcome of the auction (usually an allocation and transfers between the bidders and the auctioneer), and where $\theta_i \in \Theta_i$ and $\theta_{-i} \in \Theta_{-i}$ denote respectively the type of player i and the types of the other players. I denote the distribution of types in the support $\Theta \equiv \prod_{i=1}^n \Theta_i$ by F .

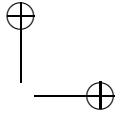
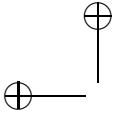
In this setting, an auction can be described by a bid space B_i per player, and a function that maps the elements of $B \equiv \prod_i B_i$ into outcomes. A strategy for Bidder i is a map $b_i : \Theta_i \rightarrow B_i$ that specifies a bid for each of Bidder i 's types.

Most of the analysis of auctions is derived under the following assumptions:

Assumption 1 (independent types across bidders) $F(\theta) = \prod_{i=1}^n F_i(\theta_i)$.

* The author gratefully acknowledges support from the Ministerio Economía y Competitividad (Spain), grants ECO2012-38863, MDM 2014-0431, and Comunidad de Madrid, MadEco-CM (S2015/HUM-3444).





Assumption 2 (quasilinear payoffs) $u_i(o, \theta) = v_i(q(o), \theta) - t_i(o)$ for $q(o)$ the allocation decision and $t_i(o)$ the transfer paid by i to the auctioneer in the outcome o .

Under these assumptions, a strategy profile (b_1^*, \dots, b_n^*) is a Bayes-Nash equilibrium if each function b_i^* maximizes Bidder i 's expected payoff conditional on her type and under the assumption that all the other bidders use b_{-i}^* :

$$b_i^*(\theta_i) \in \arg \max_{b_i \in B_i} \int_{\Theta_{-i}} (v_i(q(b_i, b_{-i}^*(\theta_{-i})), \theta) - t_i(b_i, b_{-i}^*(\theta_{-i}))) dF_{-i}(\theta_{-i}) \quad \forall i, \theta_i \in \Theta_i, \quad (12.1)$$

where I abuse the notation slightly and write q and t as functions of the bids directly rather than functions of the outcome, which in turn is a function of the bids. I denote by $U_i^{(q,t,b^*)}(\theta_i)$ the value of the maximization in (12.1). Using the quasilinearity of the bidders' payoffs, the expected revenue of the auctioneer can be written as the difference between the total expected surplus and the bidders' expected utilities:

$$\Pi^{(q,t,b^*)} \equiv \int \sum_i t_i(b^*(\theta)) dF(\theta) = \int_B \sum_i v_i(q(b^*(\theta)), \theta) dF(\theta) - \sum_i \int_{B_i} U_i^{(q,t,b^*)}(\theta_i) dF_i(\theta_i). \quad (12.2)$$

The application of the envelope theorem to (12.1) implies that the derivative of each bidder's expected payoffs in equilibrium with respect to her own type depends on the allocation function q but not on the transfer function t . This result together with the fundamental theorem of calculus and (12.2) implies one of the most remarkable results in auction theory: the revenue equivalence:

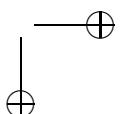
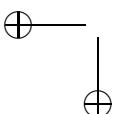
Theorem 1 Under the technical conditions required by the envelope theorem to apply to (12.1),¹ independent types and quasilinear payoffs imply that any two auction formats (q, t) and (\hat{q}, \hat{t}) with respective Bayes-Nash equilibria b^* and \hat{b}^* that implement the same allocation, i.e. $q(b^*) = \hat{q}(\hat{b}^*)$, give the same interim utility to all the bidders and the same expected revenue to the auctioneer, up to a constant, i.e. for all $i \in I$, there exists κ_i such that:

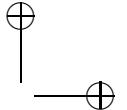
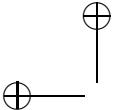
$$U_i^{(q,t,b^*)}(\theta_i) - U_i^{(\hat{q},\hat{t},\hat{b}^*)}(\theta_i) = \kappa_i, \quad \forall \theta_i \in \Theta_i$$

and

$$\Pi^{(q,t,b^*)} - \Pi^{(\hat{q},\hat{t},\hat{b}^*)} = - \sum_i \kappa_i.$$

¹ Milgrom and Segal (2002) provide a general version of the envelope theorem. Their results apply to (12.1) if Θ_i lies in a normed space and it is smoothly connected and the function $v_i(q, \theta_i, \theta_{-i})$ is differentiable in $\theta_i \in \Theta_i$ with a gradient bounded on $Q \times \Theta$, where Q is the set of all possible allocations.





Proof To prove the first display, it is sufficient to show that $U_i^{(q,t,b^*)}(\theta_i)' - U_i^{(\hat{q},\hat{t},\hat{b}^*)}(\theta_i)' = 0$. That this condition holds is a consequence of the envelope theorem since it means that:

$$U_i^{(q,t,b^*)}(\theta_i)' = \int_{\Theta_{-i}} \left(\frac{\partial v_i(q(b^*(\theta)), \theta)}{\partial \theta_i} \right) dF_{-i}(\theta_{-i}),$$

and $q(b^*) = \hat{q}(\hat{b}^*)$. The second display of the theorem follows from the first and (12.2). ■

A benign additional assumption can give us a better understanding of the power of this theorem:

Assumption 3 (a minimum type) For all i , there exists an element $\underline{\theta}_i \in \Theta_i$ such that $v_i(q, \underline{\theta}_i, \theta_{-i}) \leq v_i(q, \theta_i, \theta_{-i})$ for any $q \in Q$ and $\theta_{-i} \in \Theta_{-i}$.

Corollary 1 Any two auctions satisfying the conditions of Theorem 2 and for which the minimum type of each of the bidders exists and gets zero interim utility in a Bayes-Nash equilibrium, i.e. $U_i^{(q,t,b^*)}(\underline{\theta}_i) = U_i^{(\hat{q},\hat{t},\hat{b}^*)}(\underline{\theta}_i) = 0$, give the same interim expected payoffs to the bidders and the same expected revenue to the auctioneer.

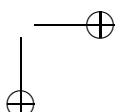
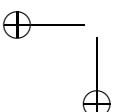
This result can be readily applied to the most basic auction formats. Think, for instance, of the first price auction and the second price auction. In both cases, one indivisible good is to be allocated, each bidder submits a price and the good is allocated to the bidder submitting the highest bid. Whereas in the first price auction, the winner pays her bid, in the second price auction, the winner pays the second highest bid. In both cases, the losers pay nothing.

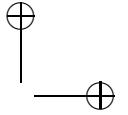
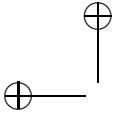
Since both the first price auction and the second price auction treat bidders symmetrically, it is natural to expect that in the symmetric case $F_i = F_j$ and $v_i = v_j$ for any i, j both auctions have an equilibrium in which all bidders use the same strictly increasing strategy. In this case, the minimum type is always outbid with probability 1, and thus gets a zero expected payoff and both auctions implement the same allocation. Consequently, if the conditions of Theorem 1 are met, both auctions give the same expected revenue to the auctioneer.

There also exists an alternative version of the revenue equivalence theorem. This is motivated by a very influential critique of game theory, in general, and auction theory in particular. The critique starts from noting that most of the applied analysis based on Bayes-Nash equilibrium usually hinges on unreasonable common knowledge assumptions about the bidders' beliefs. This critique attributed to Wilson (1987) applies to our setting in that the assumption of independent types implies that the beliefs of Bidder i about the types of the other bidders are described by F_{-i} . Since F_{-i} is common knowledge, this means that Bidder i 's beliefs about the other types are commonly known, which is odd: we assume that Bidder i has private information about her preferences, but her beliefs about the preferences of the others are common knowledge. A version of this critique applies more generally to models that do not assume independent types like Milgrom and Weber (1982).

An alternative approach to overcome Wilson's critique is to use equilibrium concepts that are agnostic with respect to the bidders' beliefs, like ex post equilibrium. In our particular model, a strategy profile $b^* = (b_1^*, \dots, b_n^*)$ is an ex post equilibrium if:

$$b_i^*(\theta_i) \in \arg \max_{b_i \in B_i} \{v_i(q(b_i, b_{-i}^*(\theta_{-i})), \theta) - t_i(b_i, b_{-i}^*(\theta_{-i}))\} \quad \forall i, \theta \in \Theta. \quad (12.3)$$





We can replicate the same steps as after (12.1) starting now from (12.3). We define the value of the maximization of problem as the ex post utility $u_i^{(q,t,b^*)}(\theta)$. We can also compute the auctioneer's revenue with a similar argument as before:

$$\pi^{(q,t,b^*)}(\theta) \equiv \sum_i t_i(b^*(\theta)) = \sum_i v_i(q(b^*(\theta)), \theta) - \sum_i u_i^{(q,t,b^*)}(\theta). \quad (12.4)$$

Finally, the envelope theorem also implies a revenue equivalence theorem (I skip the proof):

Theorem 2 *Under the technical conditions required by the envelope theorem to apply to (12.3),² quasilinear payoffs imply that any two auction formats (q, t) and (\hat{q}, \hat{t}) with respective ex post equilibria b^* and \hat{b}^* that implement the same allocation, i.e. $q(b^*) = \hat{q}(\hat{b}^*)$, give the same ex post utility to all the bidders and to the auctioneer up to a constant, i.e. for all $i \in I$, there exists $\kappa_i : \Theta_{-i} \rightarrow \mathbb{R}$ such that:*

$$u_i^{(q,t,b^*)}(\theta) - u_i^{(\hat{q},\hat{t},\hat{b}^*)}(\theta) = \kappa_i(\theta_{-i}), \forall \theta \in \Theta,$$

and

$$\pi^{(q,t,b^*)}(\theta) - \pi^{(\hat{q},\hat{t},\hat{b}^*)}(\theta) = \sum_i \kappa_i(\theta_{-i}), \forall \theta_{-i} \in \Theta_{-i}.$$

In this case, the corresponding version of Corollary 1 is stronger because the same ex post equilibrium payoffs together with the same equilibrium allocation imply that the equilibrium transfers are also the same and so the equilibrium outcome:

Corollary 2 *Any two auctions satisfying the conditions of Theorem 2 and for which the minimum type of each of the bidders gets zero utility in an ex post equilibrium, i.e. $u_i^{(q,t,b^*)}(\underline{\theta}_i, \theta_{-i}) = u_i^{(\hat{q},\hat{t},\hat{b}^*)}(\underline{\theta}_i, \theta_{-i}) = 0$, induce the same outcome for each vector of types.*

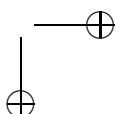
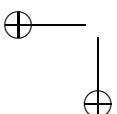
A family of auctions that have received special attention are the Vickrey-Clarke-Groves auctions (VCG in what follows). These are direct revelation auctions that implement the ex post efficient allocation $q^*(\theta) \equiv \arg \max_{q \in Q} \sum_i v_i(q, \theta)$. Direct revelation auctions are auctions in which bidders bid their types, i.e. $B = \Theta$, and in which reporting the true type is an ex post equilibrium, i.e. b^* equal to the identity is an ex post equilibrium.

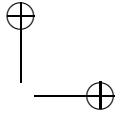
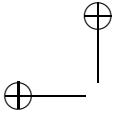
Since VCG auctions are direct revelation auctions that implement the ex post efficient allocation q^* , to complete the description of the VCG auction we only need to describe the map from reported types into net transfers to the seller that we denote by t_i^{VCG} . Since for a VCG auction,

$$u_i^{(q^*, t^{VCG}, b^*)}(\theta) = v_i(q^*(\theta), \theta_i) - t_i^{VCG}(\theta),$$

Theorem 2 implies that $t_i^{VCG}(\theta)$ is defined up to a function $\kappa_i(\theta_i)$ that depends on the other bidders' bids but not in Bidder i 's bid. Different functions $\kappa_i(\theta_i)$ give different VCG auctions.

² The same sufficient conditions as in Footnote 1 apply here.





The particular value for the transfers can be deduced from the equation above and the application of the envelope theorem, first, and the fundamental theorem of calculus, second, to compute the value of its right-hand side up to the constant $\kappa_i(\theta_i)$.

A particular VCG auction that has received special attention is the VCG auction that gives zero ex post payoffs to the minimum types $\underline{\theta}_i$ of all the bidders for all the types of the other bidders. The interest of this particular VCG auction derives from the following result:

Proposition 1 The VCG auction that gives zero ex post payoffs to the minimum types $\underline{\theta}_i$ of all bidders, i.e. $u_i^{(q^*, t^{VCG}, p^*)}(\theta) = 0$, is the VCG auction that gives maximum expected revenue among the VCG auctions that guarantee non-negative payoffs to the bidders ex post.

Proof This result follows from two facts. First, for any given allocation q , (12.4) means that the greater the bidders' expected payoffs, the lower the auctioneer's expected revenue. Second, all the types of bidder get greater payoffs in equilibrium than the minimum type. This is because the definition of the minimum type means that any other type can guarantee no less than the minimum type's utility by deviating and reporting the minimum type. ■

For simplicity, when we refer to the VCG auction we refer to the VCG of Proposition 1. As we shall see next, the transfers of the VCG auction, under the following assumption have a very intuitive interpretation:

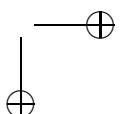
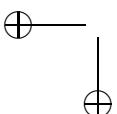
Assumption 4 (the private value assumption) $v_i(q, \theta)$ is constant in θ_{-i} , and thus we just write $v_i(q, \theta_i)$ for simplicity.

The transfers of the VCG auction under the private value assumption are:

$$t_i^{VCG}(\theta) \equiv - \sum_{j \neq i} v_j(q^*(\theta), \theta_j) + \left(\sum_{j \neq i} v_j(q^*(\underline{\theta}_i, \theta_{-i}), \theta_j) + v_i(q^*(\underline{\theta}_i, \theta_{-i}), \underline{\theta}_i) \right). \quad (12.5)$$

The first term is equal to the surplus that the other bidders derive from the implemented allocation. As we shall see next, that this surplus is transferred to the bidder and that the allocation of the VCG auction maximizes the total surplus gives incentives to the bidders to bid their true types. The second term is an adjustment term that guarantees that the minimum type $\underline{\theta}_i$ gets zero utility. Note that this second term does not depend on the bidders' report, and thus it does not affect the incentives to report. To see that these transfers together with the ex post efficient allocation q^* is a VCG auction we only need to check that truth-telling is an ex post equilibrium (which under the private value assumption is equivalent to an equilibrium in dominant strategies). This follows from the next equalities and inequality:

$$\begin{aligned} v_i(q^*(\theta), \theta_i) - t_i^{VCG}(\theta) &= \sum_j v_j(q^*(\theta), \theta_j) - \left(\sum_{j \neq i} v_j(q^*(\underline{\theta}_i, \theta_{-i}), \theta_j) + v_i(q^*(\underline{\theta}_i, \theta_{-i}), \underline{\theta}_i) \right) \\ &\geq \sum_j v_j(q^*(\tilde{\theta}_i, \theta_{-i}), \theta_j) - \left(\sum_{j \neq i} v_j(q^*(\underline{\theta}_i, \theta_{-i}), \theta_j) + v_i(q^*(\underline{\theta}_i, \theta_{-i}), \underline{\theta}_i) \right) \\ &= v_i(q^*(\tilde{\theta}_i, \theta_{-i}), \theta_i) - t_i^{VCG}(\tilde{\theta}_i, \theta_{-i}), \end{aligned}$$



where the inequality is a consequence of q^* being the ex post efficient allocation. The intuition is transparent. Since we are transferring all the surplus to the bidder, he has incentives to report truthfully so that the allocation chosen is the one that maximizes surplus.

In applications, it may be difficult to determine the minimum type of each bidder. One alternative, which I shall follow in the remainder of the chapter, is to assume that the minimum type of each bidder does not get allocated anything in the ex post efficient allocation. In this case, the transfers in (12.5) have a very remarkable interpretation when there are no externalities. Since $q^{VCG}(\underline{\theta}_i, \theta_{-i})$ does not allocate to i and there are no externalities, the second term on the right-hand side of (12.5) is equal to the maximum surplus that can be obtained by the subset of bidders that do not include i . Thus:

$$t_i^{VCG}(\theta) \equiv - \sum_{j \neq i} v_j(q^*(\theta), \theta_j) + \max_q \sum_{j \neq i} v_j(q, \theta_j). \quad (12.6)$$

Since the first term is equal to surplus obtained by the subset of bidders that do not include i , when the allocation is ex post efficient (including i), the interpretation is that i pays the “externality” that it generates on the others. Note also that in the particular case in which i does get allocated with his report, he does not pay anything as the two terms on the right-hand side of (12.6) are equal.

A particular example of a VCG auction in the private value case with no externalities is the second price auction. The winner pays the second highest bid, which is equal to the “externality” that the winner creates on the other bidders when all bidders report their true values, whereas the other bidders do not pay anything.

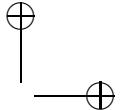
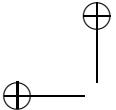
Although our analysis of Theorems 1 and 2 has been limited to static auctions, i.e. auctions in which bidders submit a single bid simultaneously and independently, it can be extended to dynamic auctions by considering their normal-form representation. The most popular dynamic auction may be the open ascending auction. In the version for one indivisible unit, the price starts at a very low level and increases continuously. All bidders start active and can declare inactive at any moment in time. This decision is irreversible. The identity of the active bidders and the price is publicly observable. The auction finishes when only one bidder remains active. This bidder gets the good and pays the last price in the auction. The other bidders do not pay anything. As we shall see in the next sections, this dynamic auction plays an important role for its properties.

In what follows, we shall assume quasilinear payoffs, private values and minimum types except when explicitly mentioned.

3 POSITION AUCTIONS IN INTERNET ADVERTISING

One of the areas in which auctions have been applied with great success in the last few years is the pricing of advertisements. This evolution has occurred in parallel with the eruption of Internet advertising, an area with revenues in the USA increasing at around 15 percent a year for the last ten years and that already accounted for \$49.5 billion in 2014.³ These online ads

³ Source: 2014 Internet Advertising Revenue Full-Year Report, accessed September 18, 2017 at <https://www.iab.com/insights/iab-internet-advertising-revenue-report-conducted-by-pricewaterhousecoopers-pwc-2/#year2014>.



are usually sold through an auction, as is the case with the big Internet companies like Google, Microsoft or Facebook.

One the most popular formats is search ad auctions. In this case, the items for sale are ads associated with keywords in Internet search engines, also called sponsored links. These ads are displayed whenever an Internet user submits a query to the search engine that matches the keyword. Since the numbers of ads that can be displayed is limited and different positions are valued differently by advertisers, it is natural to use an auction to allocate the available slots.

The auctions for search ads usually have the same basic format. Advertisers submit individual bids for different keywords. When an Internet user submits a query to the search engine, the most valuable advertisement slots are assigned to the advertisers that submitted the highest bids for the keywords that match the Internet user search. Advertisers are charged every time their ads are clicked, and hence the name “price-per-click,” an amount that depends on the auction format. Since the advertisements slots are positions in a list, these auctions are called position auctions.

The two most popular position auctions are the generalized second price auction (GSP) and a Vickrey-Clarke-Groves auction (VCG). In the former, the price-per-click of the k th most valuable advertisement slot is equal to the bid of the $k + 1$ th highest bid. In the latter, the payments are given by (12.6). Note, however, that the formula in (12.6) needs to make explicit the value functions. The most elementary version of the system used by Google and others assumes the following value functions:

Assumption 5 (the elementary click model) *Each of the slots differs from the others in the click-through rate, i.e. in the number of clicks that it produces. This click-through rate is the same across bidders and commonly known. The monetary value per click of a bidder is constant across slots and it is the bidder’s private information, i.e. the type.*

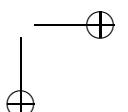
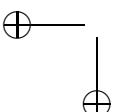
Under this assumption, the payment of Bidder i that gets slot $k \leq K$ in the VCG auction with $K < n$ slots is equal to:

$$t_i^{VCG}(\theta) = - \sum_{l \neq k} \alpha_{(l)} \theta_{(l)} + \sum_{l < k} \alpha_{(l)} \theta_{(l)} + \sum_{l > k} \alpha_{(l-1)} \theta_{(l)} = \sum_{l > k} (\alpha_{(l-1)} - \alpha_{(l)}) \theta_{(l)},$$

where $\alpha_{(k)}$ denotes the click-through rate of the slot $k \leq K$ under the convention that slots are ordered according to their click-through rates. For notational convenience, we let $\alpha_{(k)} \equiv 0$ for any $k > K$.

If $K > 1$, these payments depend on the types of more than one of the other bidders. However, the price that each bidder pays in the GSP auction is the bid of another bidder. Hence, this price can only depend on at most one type of the other bidders when the bidder’s type is her private information. Thus, one can deduce from the ex post equivalence implied by Theorem 2 that there is no ex post equilibrium of the GSP auction that implements the ex post efficient allocation. Gomes and Sweeney (2014) study the set of Bayes-Nash equilibria of the GSP.

In reality, however, the search ad auctions occur so frequently that one can imagine that at some point competing bidders might have been able to figure out the types of the others. Accordingly, people have argued that it may be reasonable to analyze the equilibrium assuming that types are commonly known. In this case, Varian (2007) and Edelman,



Ostrovsky, and Schwarz (2007) show that the GSP auction has an equilibrium in which the bidder with the $k + 1$ th highest-type $\theta_{(k+1)}$ bids:

$$\sum_{l>k} (\alpha_{(l-1)} - \alpha_{(l)})\theta_{(l)}, \quad (12.7)$$

where α_0 can be any value strictly larger than α_1 . It is easy to see that this equilibrium is outcome equivalent to the truthful equilibrium of the VCG auction. Varian (2007) and Edelman et al. (2007) also show that the GSP auction with common knowledge of types has other equilibria. Any of these equilibria that satisfies a natural additional condition gives weakly greater revenue than the VCG auction.

One flaw of the previous argument is that bidders may have incentives to distort their early bids to mislead the other bidders' beliefs, and this can preclude the learning of types in equilibrium. A full analysis of this argument seems a complex exercise as one has to model a dynamic model in which bidders' beliefs evolve according to the observed behavior of the rivals. One alternative and simpler approach is to assume that bidders may have a direct communication channel through which they can disclose their types if they find it convenient. One can model this approach with the following two-stage game:

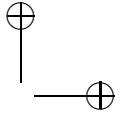
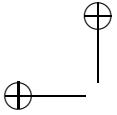
Definition 1 We call the CT+GSP auction a two-stage auction in which bidders make a cheap talk (CT) announcement in the first stage and participate in a GSP auction in the second stage.

If this auction has no equilibrium with truthful reporting of types in the first stage, it is reasonable to conjecture that there is no equilibrium of an infinitely repeated GSP auction in which bidders' types get revealed in finite time as the discount rate tends to 1. The basis for this claim is that it is meaningful to approximate the limit of the equilibrium when the discount factor goes to 1 by the limit of the equilibrium of the time average payoffs. Hence, the payoffs in any finite sequence of auctions have no effect on the time average payoffs and thus can be identified with the cheap talk stage of the CT+GSP auction.

Regretfully, there is no truthful reporting equilibrium of the CT+GSP auction, in general, that implements the ex post efficient allocation in an ex post equilibrium. To see why, note that Corollary 2 implies that such equilibrium must have bids in the second stage as in (12.7) but with the other bidders' types replaced by the reported types in the cheap talk stage. This means that in the case of three bidders and two slots with α_1 arbitrarily close to α_2 , the bidder with the second highest reported type bids arbitrarily close to $\alpha_2\hat{\theta}_{(3)}$. Thus, the bidder with lowest type can do strictly better by deviating and reporting in the cheap talk stage $\hat{\theta}_{(3)} = 0$ and bidding slightly above zero in the auction. In this case, the bidder gets allocated the second slot at a price arbitrarily close to zero.

A number of subsequent theoretical papers have studied extensions of the elementary click model. For instance, Edelman and Schwarz (2010) study the introduction of optimal reserve prices. Most of the subsequent theoretical literature has focused on the implication of withdrawing the assumption of fixed click-through rates constant across bidders.

Liu and Chen (2006) and Liu, Chen, and Whinston (2010) notice that in reality click-through rates differ across bidders and study an auction design implemented by Google in which the bids of the bidders are weighted by a measure of the bidder click-through rate inferred from past performance. They show that efficiency can be implemented with weights



equal to the true click-through rates and revenue maximization requires weights biased in favor of bidders with lower click-through rates.

Athey and Ellison (2011) study a model in which the ad position may be a useful signaling device to convey the quality of the advertiser measured by the probability of satisfying the searcher's needs. In this model, clicking rates are endogenously determined by the searchers' optimal clicking strategies. This is also a feature of Gomes (2014). The difference from Athey and Ellison's model is that in Gomes (2014) the most relevant advertisers do not coincide with those with highest willingness to pay per click, making the optimal auction design non-standard.

Aggarwal, Goel, and Motwani (2006) and Jeziorski and Segal (2015) analyze and document that the click-through rate of an advertisement depends on the other advertisements simultaneously displayed and on their positions. This affects the design of the revenue-maximizing mechanism and the mechanisms that implement the ex post efficient allocation.

Börgers et al. (2013) use a parsimonious theoretical model to test whether the observed bids in position auctions can be rationalized as a Nash equilibrium. They find that this is the case only for relatively short periods of time. For longer periods of time one has to allow for an unexplained structural change in the preferences. They also find that an advertiser's value per click decreases in the listing position.

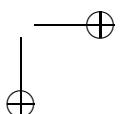
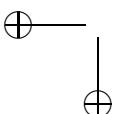
Yang and Ghose (2010) and Blake, Nosko, and Tadelis (2015) study the effectiveness of paid search ads in the presence of organic listings, i.e. the list of websites that the Internet search engine provides according to their ranking algorithm. In this case, consumers face two competing lists of results that may be relevant to their search: (i) the sponsored search listing and (ii) the organic search listing. Yang and Ghose's 2010 empirical finding is that click-through rates in organic listings have a positive interdependence with click-through rates in sponsor search advertising. However, Blake, Nosko, and Tadelis (2015) show that the effectiveness of search advertisement is small for ads of well-known brands or for informed consumers to the extent of giving negative average returns.

Other issues that remain to be explored are competition for advertisers across search platforms and the interaction of search ad auctions and non-search advertisements and their optimal design. There also seems to exist some role for richer models of uncertainty and search in the literature of position auctions.

4 INTERNET AUCTIONS

In the last few years, auction theory has also been developed extensively to cope with another area of innovation in the market place, the rise of auctions sites in which both professional and occasional sellers put up for sale items to potential buyers. The most popular is eBay with total revenues of \$8.59 billion in 2015. Initially, most of the items were second-hand goods, but this has been changing. Recently, slightly more than half of the items for sale are new.

The basic eBay auction is a variation of the open ascending auction closer to the traditional English auction. The price starts at an initial price (called "first bid") fixed by the seller. At any time a bidder can submit a bid at a minimum increment greater than the current price. The current price increases to this bid and the bidder becomes the current winner. The auction ends at a predetermined moment, usually between three to ten days after opening. The current winner wins the item and pays the current price if the price is above a reserve price (called "reservation price") secretly fixed by the seller.



The basic eBay auction allows for two additional features: a buy-it-now price and proxy bidding. The buy-it-now price offers the possibility to bidders of buying the item at a price predetermined by the seller. In the usual setting, the buy-it-now price disappears once a bidder submits a bid above the reserve price. The proxy bidding allows the bidder to instruct eBay to outbid any bid up to a maximum level. The purpose of proxy bidding is to alleviate the burden of keeping track of the evolution of prices in the auction.

The buy-it-now price and the proxy bidding put the basic eBay auction in between two standard selling mechanisms. The former makes the auction a traditional posted price sale until a bid is submitted. Indeed, by fixing the buy-it-now price equal to the initial bid the auction becomes almost equivalent to a posted price. The proxy bidding brings the auction closer to the standard VCG auction corresponding to the private value assumption. Indeed, if only proxy bids are allowed and bidders cannot submit more than one proxy bid, the auction becomes a second price auction, which is the VCG auction that corresponds to private values.

eBay charges a fee to the seller that depends on the first bid, the reservation price and the final price. eBay also has a mechanism to keep track of the reputation of buyers and sellers. After each transaction, the seller and the buyers report whether the transaction was positive, neutral or negative, giving respective values +1, 0, and -1. A more detailed report is also available for buyers. eBay makes serious effort into providing useful summaries of the feedback received by each buyer or seller.

The practice of eBay and related Internet auctions have opened several theoretical and empirical issues, most of the time interrelated. Here, I shall follow the approach of Hasker and Sickles (2010) and distinguish sellers' related issues and buyers' related issues. I start with the former.

4.1 Sellers' Related Issues

4.1.1 Buy-it-now prices

From the point of view of traditional auction theory the use of buy-it-now prices is surprising as it does not seem consistent with the general solution for optimal auctions proposed by Myerson (1981). However, as Bose and Daripa (2009) note, there are some particular distribution functions that do not satisfy Myerson's 1981 regularity assumption for which an auction with a buy-it-now price is optimal.

Alternatively, Reynolds and Wooders (2009) show that a buy-it-now price is revenue enhancing in the case in which the seller is risk neutral and the bidders risk averse. Their results also show that a "permanent" buy-it-now price (used in Yahoo auctions) that remains available along the auction gives greater expected revenue than the "temporary" version used in eBay. Mathews and Katzman (2006) show that the use of a buy-it-now price is optimal for a risk-averse seller that sells to risk-neutral bidders. Gallien and Gupta (2007) propose another explanation: that the bidders and the seller may be impatient.

Shahriar and Wooders (2011) provide experimental evidence that a "temporary" buy-it-now price raises the seller's expected revenue in the private value case.

4.1.2 Asymmetric information, fraud and feedback

The lack of physical contact in eBay sales create a double asymmetric information problem: buyers cannot be certain whether the quality of the good delivered is as promised by

the seller and whether the seller will default in his obligation by either not delivering the good or delivering a counterfeit item. eBay uses the feedback mechanism that I described above to minimize these two problems. Several empirical studies have tried to assess how relevant these two problems are, and to what extent the eBay feedback system solves them.

Dewan and Hsu (2004), Eaton (2007), Houser and Wooders (2006), Livingston (2005), Lucking-Reiley et al. (2007), McDonald and Slawson (2002), Melnik and Alm (2002) and Resnick and Zeckhauser (2002) point out that there exists a negative correlation between prices and feedback. This suggests that asymmetric information is perceived as a problem by bidders.

Cabral and Hortacsu (2010) also provide empirical evidence that suggests that auctioneers may build a good reputation accumulating positive feedback in perhaps small transactions to later profit from default in more profitable transactions.

Jin and Kato (2006) use a field experiment to test directly for the honesty of sellers. They find that sellers making the boldest claims with respect to the quality of their products are more likely to commit fraud. They also find that sellers with a high reputation on eBay are less likely to make bold claims and less likely to commit fraud. However, conditional on no fraud, the actual quality of the item is uncorrelated with seller ratings. Their results also suggest that unexperienced bidders tend to fall prey to defaulting sellers.

4.1.3 Shill bidding and secret reserve prices

The eBay rules allow a seller to fix a secret reserve price. Besides, the seller may submit a shill (fake) bid to increase the final price. Both policies induce a minimum price in the auction that is not observable by the bidders. The difference, though, is that shill bids allow the seller to update this minimum price as the auction evolves. Besides, shill bids may alter the bidders' beliefs about a common component of the value of the good for sale.

The analysis of Myerson (1981) implies that none of these strategies is revenue increasing if it is anticipated by the bidders and the auctioneer can use a public reserve price (like the first bid in eBay). Indeed, Lamy (2009) shows that a seller unable to commit not to fix a secret reserve price in an open ascending auction may prefer to use a first price auction instead. This is because a secret reserve price may reduce the seller's expected revenue and a secret reserve price has no role in a first price auction. However, Graham, Marshall, and Richard (1990) provide a setting with heterogeneous bidders in which shill bidding is optimal as it implements a history-dependent reserve price. Finally, Jehiel and Lamy (2015a) show that secret reserve prices and public reserve prices can coexist in a market with many heterogeneous sellers if there are some bidders with naive beliefs about the distribution of the secret reserve prices.

Allowing for common values introduces new effects. In particular, both shill bids or secret reserve prices have the advantage over public reserve prices that may induce greater entry, which may provide useful information to the bidders. Vincent (1995) shows that this effect is sufficient to increase the seller's revenue by application of the linkage principle; see also Milgrom (1985). However, Chakraborty and Kosmopoulou (2004) show that the seller's profits can go down with shill bids if the seller places some value on retaining the good.

In the particular case of eBay auctions, Engelberg and Williams (2009) suggest that an auctioneer can profit from shill bidding because bidders tend to submit rounded numbers and thus there is little risk in increasing the prices up to the next rounded number. The profitability

of this strategy is amplified by the details of eBay's rule as it shows an example provided by Engelberg and Williams (2009, p. 511), which I reproduce here

[Suppose] the starting price is \$1.50, and the bid increment is \$0.50. Assume for the moment that bidders only bid on the dollar. At time 1, bidder B places a bid of \$4.00, and the price remains at \$1.50. Suppose the seller, S, has a shill account and desires to increase the price as much as possible without becoming the high bidder. At time 2, S places a bid of \$2.00, after which the price goes to \$2.50. Since bidders are assumed to only bid on the dollar, S knows that B's bid is at least \$3.00. At time 3, S bids \$3.00 and the price moves to \$3.50, so at time 4 S bids \$4.00, and the price only moves to \$4.00. S then realizes that \$4.00 is B's maximum bid, so S cannot push the price higher without becoming the high bidder, and he stops bidding.

Kauffman and Wood (2005) also provide evidence of shill bidding by sellers that want to implement a reserve price but avoid eBay's fees on the first bid or the reservation price.

4.2 Buyers' Related Issues

4.2.1 Incremental bidding

One documented phenomenon is that bidders usually revise their bids several times. Wilcox (2000), for instance, shows that a bidder submits on average between 1.5 and two bids, and Ockenfels and Roth (2006) report that 38 percent of bidders in their data sample bid at least twice. This is surprising in a private value setting as the bidder has the less time-consuming option of submitting a proxy bid equal to her true value. Bajari and Hortacsu (2003) and Ockenfels and Roth (2006) provide a model with common value components in which bidders only submit bids once and just before the end of the auction. However, Peters and Severinov (2006) show that incremental bidding is the equilibrium outcome of a decentralized model of trade in which bidders can submit bids in several simultaneous auctions.

4.2.2 Sniping

Another documented, known as sniping, is that a large fraction of bidding activity occurs close to the end of the auction, so close that some of the bids are rejected because they arrive too late. This is surprising as a bidder can always guarantee winning whenever it is profitable by submitting an earlier proxy bid equal to her value, at least in the private value case. A possible explanation is to be found in the theoretical models of Bajari and Hortacsu (2003) and Ockenfels and Roth (2006) that we already mentioned above. They show that common value components can make better-informed bidders submit bids just before the end of the auction to avoid the free-riding of their information. This seems an appealing argument since bidders are indifferent with respect to the timing of their bid submission in a pure private value model with no discounting and no late bids rejected. Another explanation put forward by Ockenfels and Roth (2006) is that late bidding may be a best response to incremental bidding: a bidder that bids late leaves no time for the bid escalation of the rivals. Another explanation is Roth and Ockenfels' (2002) "snipe or war" strategy. According to this strategy, any early bid is contested by a price war that leads to very competitive prices. Thus, bidders may find it beneficial to submit a late bid and win with certain probability rather than bidding early and entering the "war phase." Hasker and Sickles (2010) also suggest that sniping may also be a best response to shill bidding.

In all these theoretical models, sniping reduces the seller's revenue. This seems to be consistent with some empirical evidence, albeit weak. Ely and Hossain (2009) find that

bidding in the last five seconds of the auction gives 1 percent more surplus to the bidders when compared to bidding early in the auction and only once. However, a field experiment by Gray and Reiley (2013) shows no statistically significant effect of late bidding on the bidders' surplus.

4.2.3 Over-bidding

Malmendier and Lee (2011) report that in a representative example of eBay online auctions, the final prices were higher than the buy-it-now price available simultaneously on eBay on 42 percent of the cases, and if shipping costs are included the number increases by up to 73 percent. They also show that once shipping costs are included, the expected auction price is higher than the fixed price at a statistically significant level. They suggest that the only possible explanation is either limited attention or that bidders derive utility from winning an auction. Malmendier and Lee (2011) are inclined in favor of the former as the latter seems ad hoc and hard to falsify. Schneider (2015), however, argues that search costs can also explain this evidence.

5 COMBINATORIAL AUCTIONS, VCG AUCTIONS AND THE CORE

A natural extension of auction theory that has been studied in the last few years is the design of combinatorial auctions. These are auctions in which bidders bid for packages of items rather than for individual items. These auctions make sense when a bidder's willingness to pay for the package is not equal to the sum of the willingness to pay for each individual item in the package. An early real-life example arose in the spectrum auctions that were conducted in the 1990s: companies bidding for rights to transmit signals over specific bands of the electromagnetic spectrum noted that their willingness to pay for a license on a given frequency in a given area depended on whether they acquire a similar license in either the same area (they may be substitutes) or an adjacent area (they may be complements). Other examples are the sale of airport slots for take-off and landing, or auctions for bus routes.

The design of these auctions has brought new challenges that require techniques that go beyond economics and indeed many of the most recent advances come from operations research and computer science. The need for these techniques is justified by the complexity inherent in the description of bids and the selection of the allocation as the number of items grows large. Think, for instance, about the airport slots allocation problem. Both the number of airports and the number of slots per airport is relatively large, and thus the vector of valuations and bids that the bidders and the auctioneer, respectively, must evaluate. Cramton, Shoham, and Steinberg 2006 provide a very good introduction to the recent advances in these areas. Here, I shall focus only on the economic analysis.

If we leave aside complexity issues, the natural candidate for a combinatorial auction is the corresponding VCG auction. There are three reasons. First, it seems difficult to beat in terms of revenue for the auctioneer if one is interested in implementing the ex post efficient allocation with a mechanism in which the bidders that get nothing pay nothing; see the revenue equivalence in Corollaries 1 and 2. Second, the strategic problem for the bidders is relatively simple: they only need to report their true preferences for the allocations. Third, it can be easily modified to encompass any additional restriction on the set of feasible allocations like

a minimum number of items allocated to certain group of bidders, or a limitation on the concentration of the units allocated among the bidders.

In reality, however, applying the VCG auction to a combinatorial setting present several drawbacks. Ausubel and Milgrom 2006 point out that in the case that at least one bidder's preferences displays complementarities the following outcomes may occur:

- (a) Low revenues.
- (b) Revenues may decrease when either the bids or the number of bidders increase.
- (c) A winning bidder can reduce the price paid by pretending to be several different bidders.
- (d) Efficient mergers between bidders may be discouraged.
- (e) A losing bidder can manipulate the auction and win at a profit by pretending to be several different bidders.
- (f) The losing bidders could collude and win, instead, at low prices.

None of them happens when all the bidders' preferences only display substitutabilities. These results are extensively analyzed by Ausubel and Milgrom (2006). Here, I will only reproduce a variation of their examples for the sake of illustration.

To illustrate (a) suppose the sale of two licenses to operate the radio spectrum and three bidders. Bidder 1 is only interested in both licenses together and his willingness to pay for this package is \$2 billion. Bidders 2 and 3 are interested in only one license and each is willing to pay for the license \$2 billion plus $\epsilon > 0$. In this case, selling an indivisible package that includes both licenses through an auction (e.g. a second price auction) reports a revenue of \$2 billion plus ϵ whereas the VCG auction that allows all the combinatorial bids gives zero revenue: it allocates efficiently one license to Bidder 2 and one to Bidder 3 but charges them a price of zero. The latter can be deduced from the fact that the VCG charges each bidder the externality induced on the others. For instance, the sum of payoffs of Bidder 1, Bidder 2 and the auctioneer is equal to \$2 billion plus ϵ both in the case in which Bidder 3 participates and when he stays out of the auction. Thus, Bidder 3's externality is zero and thus pays nothing in the VCG auction.

The same example also illustrates (b) and (c). Suppose that either Bidder 3 is absent or his value (and bid) is reduced to zero. In this case, one of the licenses is allocated to Bidder 2 at a price equal to the externality on Bidder 1, i.e. \$2 billion. This result is not only counterintuitive (and probably hard to justify to the public) but also implies drawback (c). An example of (c) is that Bidder 2 could reduce the price that she pays by pretending that there is an additional bidder that puts value \$2 billion plus ϵ in one single license. Of course, this is only feasible if the auctioneer cannot track the real identity of the bidders.

A variation of the arguments in previous paragraph also illustrates (d). Suppose in our original example a possible merger before the auction between Bidders 2 and 3 with synergies of $Y > 0$ billion dollars. This merger is clearly socially desirable. However, if both bidders merge they can anticipate that although they are set to win in the VCG auction, the price that they will pay is \$2 billion, i.e. the externality on Bidder 1. Thus, the merger will be undertaken only if the synergies Y are sufficiently large to compensate for the price increase from zero to \$2 billion. Thus mergers with synergies Y between zero and \$2 billion are socially desirable but are not undertaken when a VCG auction is used.

To illustrate (e) suppose a variation of the original example in which Bidder 2 and Bidder 3 are replaced by another bidder, say Bidder 4, that puts a value of \$0.5 billion on a single

license and a value of \$1 billion on both licenses. In a VCG auction this bidder loses against Bidder 1. However, this bidder can win both licenses at a price of zero by pretending to be two bidders identical to the Bidders 2 and 3 of the original example.

Finally, to illustrate (f) one can use another variation of the original example in which Bidder 2 and Bidder 3 have values of \$0.5 billion each. In this case, the VCG does not allocate the good to them, but they could collude and instead bid as in the original example, which guarantees them a unit for each at a price of zero.

The versions of the revenue equivalence theorem in Theorems 1 and 2 suggests that there is very little that can be done if one finds it reasonable to be restricted to the implementation of the ex post efficient allocation. In reality, however, one may be happy to sacrifice some ex post efficiency to avoid some of the previous drawbacks and in particular to boost revenues. Alternatively, there may be other auction formats that still implement the ex post efficient allocation without any of the previous drawbacks when the bidders' preferences are commonly known. These approaches have inspired new auction designs like the ascending proxy auction (see Ausubel and Milgrom, 2002), the simultaneous ascending auction (see Cramton, 1998 and Milgrom, 2000), and the clock-proxy auction (Ausubel, Cramton, and Milgrom, 2006).

6 OTHER TOPICS

I conclude this survey by summarizing a list of recent contributions organized by topics.

6.1 Auctions and Resale

In many realistic environments, bidders have the option of trading among them in a secondary market after the auction. Garratt and Tröger (2006) show that under the usual private value assumption, a bidder that places no value on the good can win the auction in equilibrium for the sole purpose of reselling it in the secondary market. Hafalir and Krishna (2008) extend this analysis to the case of general heterogeneous bidders with possibly different type distributions. They show that in this case the first price auction gives larger expected revenue than the second price auction. This result contrasts sharply with the case in which there is no secondary market where Maskin and Riley (2000) show that there is no general revenue ranking.

6.2 Auctions and Entry

As Milgrom (2004, p. 247) puts it, "many of the most important practical issues in auction design concern the interaction of the design and entry decisions." A sequence of classical papers starting with McAfee (1993) and Levin and Smith (1994) have noted for the private value model that costly entry implies that for the case of ex ante symmetric bidders any ex post efficient mechanism maximizes the seller's expected payoffs. Jehiel and Lamy (2015b) note how this result applies to the case of asymmetric bidders, whereas Moreno and Wooders (2011) show that the result for ex ante symmetric bidders hinges on the maintained assumption that all bidders have the same entry cost. Interestingly, Gorkem and Okan (2009) note that in the case of ex ante symmetric bidders, the seller's maximum payoffs may be achieved with either an auction that treats bidders asymmetrically or in an asymmetric equilibrium of an auction that treats bidders symmetrically.

In a related work, Bulow and Klemperer (2009) study how costly entry affects the ranking of two common mechanisms for the sale of companies: a sequential sale in which the seller sequentially receives offers from potential buyers, and a standard auction in which all potential buyers submit a bid simultaneously. Their main result shows that sequential sales are in general more efficient than auctions, but the seller may prefer an auction. This explains some empirical regularities observed in the sale of companies.

6.3 Financial Constraints in Auctions

That financial or budget constraints may interfere with the properties of auctions has been known since the pioneering studies of Pitchik and Schotter (1988), Che and Gale (1998) and Maskin (2000). Subsequent work has explored the effect of financial constraints in single-unit auctions: Fang and Parreiras (2002, 2003), Malakhov and Vohra (2008) and Pitchik (2009), and in multi-unit auctions, Benoit and Krishna (2001), Brusco and Lopomo (2008, 2009), Ashlagi et al. (2010), Hafalir, Ravi, and Sayedi (2012), and Dobzinski, Lavi, and Nisan (2012).

However, the prevalent view was that “auctions [still] work well if raising cash for bids is easy” (Aghion, Hart, and Moore, 1992, p. 527). More recently, this view has been contested by Rhodes-Kropf and Viswanathan (2005) and by Beker and Hernando-Veciana (2015). The former paper considers the case in which firms finance their bids in a competitive financial market and the latter the case in which firms finance their bids through retained earnings. Burkett (2015, 2016) also studies auctions with endogenous budgets, but in his models they are chosen by a third party.

Financial constraints also create the possibility of defaults, which can give rise to undesired bidder behavior. Zheng (2001) and Calveras, Ganuza, and Hauk (2004) show that the bidders with greater probability of defaulting may bid higher, and thus win with greater probability than bidders with lower probability of defaulting.

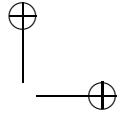
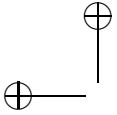
Finally, DeMarzo, Kremer, and Skrzypacz (2005) and Yeon-Koo Che (2010) explore the consequences of using financial instruments rather than money to pay the bid.

6.4 A General Type Space

As I explain in Section 2, one of the weak points of auction theory, summarized in Wilson’s critique, is that most of its analysis is done under unrealistic assumptions on the common knowledge of beliefs. Regretfully, very little has been done in this direction except the study of ex post equilibria, which turns out to be a too demanding equilibrium concept: many auction games, like the first price auction, have no ex post equilibrium in general. An exception is a recent paper by Bergemann, Brooks, and Morris (2015) that characterizes the lowest winning bid distribution that can arise across all information structures and equilibria associated with a given symmetric prior distribution over values.

6.5 Open Ascending Auctions and Efficiency

In the private value case, I already noted that the efficient allocation can be implemented using a VCG auction, for instance, a second price auction in the case of single unit sales. The problem, however, becomes more complex when the value functions have common value components. Of course, one can appeal to the revelation principle and show that if the ex post



334 *Handbook of game theory and industrial organization: applications*

efficient allocation can be implemented with an auction, then it must be the case that it can be implemented with a direct revelation auction. This simple idea can allow us to construct auctions that have been interpreted very much the same way as the VCG auctions.

In reality, however, VCG auctions for common values are less appealing as their payoff structure is less straightforward. Instead, the literature, in particular Birulin and Izmalkov (2011) and Dubra, Echenique, and Manelli (2009), has focused on understanding the conditions under which the ex post efficient allocation can be implemented with an open ascending auction. This is a realistic mechanism that combines two appealing ideas. On the one hand, it is a competitive mechanism in the sense that the bidder cannot affect the price at which she buys, much the same as a second price auction. On the other hand, it allows for a lot of information disclosure during the auction that may be helpful for the bidders to discover their true values.

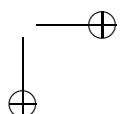
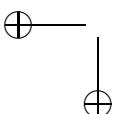
One difference with the private value case, though, is that there are many natural examples in which there is no auction that implements the ex post efficient allocation. (Hernando-Veciana and Michelucci, 2011, p. 496) provide the following example:

Suppose that an oil tract is put up for sale between two wildcatters. The first one, the incumbent, has a high marginal cost and a low fixed cost, whereas the second one, the entrant, has a low marginal cost and a high fixed cost. In this case, it may be efficient to allocate the good to the incumbent if there is little oil and to the entrant if there is much oil. However, Maskin (1992) has shown that this allocation, i.e. the first best, is not implementable when the amount of oil is private information of the incumbent.

In examples like this, one can define the second best as the allocation that maximizes the expected social surplus, i.e. the difference between the buyer's value and the seller's cost between the agents that trade, among the allocations that are implementable. Under certain conditions, which include the example above, the second best can be implemented with an open ascending auction if there are two bidders (see Hernando-Veciana and Michelucci, 2011) but not if there are more than two bidders (Hernando-Veciana and Michelucci, 2013).

REFERENCES

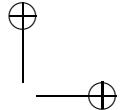
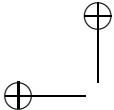
- Aggarwal, G., A. Goel, and R. Motwani (2006): "Truthful Auctions for Pricing Search Keywords," in *Proceedings of the 7th ACM Conference on Electronic Commerce*, pp. 1–7.
- Aghion, P., O. Hart, and J. Moore (1992): "The Economics of Bankruptcy Reform," *Journal of Law, Economics and Organization*, 8(3), 523–546.
- Ashlagi, I., M. Braverman, and A. Hassidim, R. Lavi, and M. Tennenholtz (2010): "Position Auctions with Budgets: Existence and Uniqueness," *The BE Journal of Theoretical Economics*, 10(1).
- Athey, S., and G. Ellison (2011): "Position Auctions with Consumer Search," *The Quarterly Journal of Economics*, 126, 1213–1270.
- Ausubel, L.M., and P.R. Milgrom (2002): "Ascending Auctions with Package Bidding," *Advances in Theoretical Economics*, 1(1), 1–42.
- Ausubel, L.M., and P.R. Milgrom (2006): The Lovely but Lonely Vickrey Auction, Chapter 2 in P.C. Cramton, Y. Shoham, and R. Steinberg (eds), *Combinatorial Auctions*. Cambridge, MA: MIT Press.
- Ausubel, L.M., P.C. Cramton, and P. Milgrom (2006), Chapter 5 in P.C. Cramton, Y. Shoham, and R. Steinberg (eds), *Combinatorial Auctions*. Cambridge, MA: MIT Press.
- Bajari, P., and A. Hortacsu (2003): "The Winner's Curse, Reserve Prices, and Endogenous Entry: Empirical Insights from eBay Auctions," *RAND Journal of Economics*, 34(2), 329–355.
- Beker, P.F., and Á. Hernando-Veciana (2015): "The Dynamics of Bidding Markets with Financial Constraints," *Journal of Economic Theory*, 155, 234–261.



- Benoit, J.-P., and V. Krishna (2001): "Multiple-object Auctions with Budget-constrained Bidders," *The Review of Economic Studies*, 68(1), 155–179.
- Bergemann, D., B. Brooks, and S. Morris (2015): "First Price Auctions with General Information Structures: Implications for Bidding and Revenue," *Cowles Foundation Discussion Papers* 2018R, Cowles Foundation for Research in Economics, Yale University.
- Binmore, K., and P. Klemperer (2002): "The Biggest Auction Ever: The Sale of the British 3G Telecom Licences," *The Economic Journal*, 112(478), C74–C96.
- Birulin, O., and S. Izmalkov (2011): "On Efficiency of the English Auction," *Journal of Economic Theory*, 146(4), 1398–1417.
- Blake, T., C. Nosko, and S. Tadelis (2015): "Consumer Heterogeneity and Paid Search Effectiveness: A Large-scale Field Experiment," *Econometrica*, 83(1), 155–174.
- Börgers, T., I. Cox, M. Pesendorfer, and V. Petricek (2013): "Equilibrium Bids in Sponsored Search Auctions: Theory and Evidence," *American Economic Journal: Microeconomics*, 5(4), 163–187.
- Bose, S., and A. Daripa (2009): "Optimal Sale Across Venues and Auctions with a Buy-now Option," *Economic Theory*, 38(1), 137–168.
- Brusco, S., and G. Lopomo (2008): "Budget Constraints and Demand Reduction in Simultaneous Ascending-bid Auctions," *The Journal of Industrial Economics*, 56(1), 113–142.
- Brusco, S., and G. Lopomo (2009): "Simultaneous Ascending Auctions with Complementarities and Known Budget Constraints," *Economic Theory*, 38(1), 105–124.
- Bulow, J., and P. Klemperer (2009): "Why Do Sellers (Usually) Prefer Auctions?" *The American Economic Review*, 99(4), 1544–1575.
- Burkett, J. (2015): "Endogenous Budget Constraints in Auctions," *Journal of Economic Theory*, 158, 1–20.
- Burkett, J. (2016): "Optimally Constraining a Bidder Using a Simple Budget," *Theoretical Economics*, 11, 133–155.
- Cabral, L., and A. Hortacsu (2010): "The Dynamics of Seller Reputation: Evidence from eBay," *The Journal of Industrial Economics*, 58(1), 54–78.
- Calveras, A., J.-J. Ganuza, and E. Hauk (2004): "Wild Bids. Gambling for Resurrection in Procurement Contracts," *Journal of Regulatory Economics*, 26(1), 41–68.
- Chakraborty, I., and G. Kosmopoulos (2004): "Auctions with Shill Bidding," *Economic Theory*, 24(2), 271–287.
- Che, Y.-K., and I. Gale (1998): "Standard Auctions with Financially Constrained Bidders," *The Review of Economic Studies*, 65(1), 1–21.
- Cramton, P.C. (1995): "Money Out of Thin Air: The Nationwide Narrowband PCS Auction," *Journal of Economics & Management Strategy*, 4(2), 267–343.
- Cramton, P. (1998): "Ascending Auctions," *European Economic Review*, 42(3), 745–756.
- Cramton, P.C., Y. Shoham, and R. Steinberg (eds) (2006): *Combinatorial Auctions*. Cambridge, MA: MIT Press.
- DeMarzo, P.M., I. Kremer, and A. Skrzypacz (2005): "Bidding with Securities: Auctions and Security Design," *The American Economic Review*, 95(4), 936–959.
- Dewan, S., and V. Hsu (2004): "Adverse Selection in Electronic Markets: Evidence from Online Stamp Auctions," *The Journal of Industrial Economics*, 52(4), 497–516.
- Dobzinski, S., R. Lavi, and N. Nisan (2012): "Multi-unit Auctions with Budget Limits," *Games and Economic Behavior*, 74(2), 486–503.
- Dubra, J., F. Echenique, and A.M. Manelli (2009): "English Auctions and the Stolper-Samuelson Theorem," *Journal of Economic Theory*, 144(2), 825–849.
- Eaton, D.H. (2007): "The Impact of Reputation Timing and Source on Auction Outcomes," *The BE Journal of Economic Analysis & Policy*, 7(1).
- Edelman, B., and M. Schwarz (2010): "Optimal Auction Design and Equilibrium Selection in Sponsored Search Auctions," *The American Economic Review*, 100(2), 597–602.
- Edelman, B., M. Ostrovsky, and M. Schwarz (2007): "Internet Advertising and the Generalized Second-price Auction: Selling Billions of Dollars Worth of Keywords," *American Economic Review*, 97(1), 242–259.
- Ely, J.C., and T. Hossain (2009): "Sniping and Squatting in Auction Markets," *American Economic Journal: Microeconomics*, 1(2), 68–94.
- Engelberg, J., and J. Williams (2009): "eBay's Proxy Bidding: A License to Shill," *Journal of Economic Behavior & Organization*, 72(1), 509–526.
- Fang, H., and S.O. Parreiras (2002): "Equilibrium of Affiliated Value Second Price Auctions with Financially Constrained Bidders: The Two-bidder Case," *Games and Economic Behavior*, 39(2), 215–236.
- Fang, H., and S.O. Parreiras (2003): "On the Failure of the Linkage Principle with Financially Constrained Bidders," *Journal of Economic Theory*, 110(2), 374–392.
- Gallien, J., and S. Gupta (2007): "Temporary and Permanent Buyout Prices in Online Auctions," *Management Science*, 53(5), 814–833.
- Garratt, R., and T. Tröger (2006): "Speculation in Standard Auctions with Resale," *Econometrica*, 74(3), 753–769.
- Gomes, R. (2014): "Optimal Auction Design in Two-sided Markets," *The RAND Journal of Economics*, 45(2), 248–272.

- Gomes, R., and K. Sweeney (2014): "Bayes-Nash Equilibria of the Generalized Second-price Auction," *Games and Economic Behavior*, 86, 421–437.
- Gorkem, C., and Y. Okan (2009): "Optimal Auctions with Simultaneous and Costly Participation," *The B.E. Journal of Theoretical Economics*, 9(1), 1–33.
- Graham, D.A., R.C. Marshall, and J.-F. Richard (1990): "Phantom Bidding Against Heterogeneous Bidders," *Economics Letters*, 32(1), 13–17.
- Gray, S., and D.H. Reiley (2013): "Measuring the Benefits to Sniping on eBay: Evidence from a Field Experiment," *Journal of Economics and Management*, 9(2), 137–152.
- Hafalir, I., and V. Krishna (2008): "Asymmetric Auctions with Resale," *American Economic Review*, 98(1), 87–112.
- Hafalir, I.E., R. Ravi, and A. Sayedi (2012): "A Near Pareto Optimal Auction with Budget Constraints," *Games and Economic Behavior*, 74(2), 699–708.
- Hasker, K., and R. Sickles (2010): "eBay in the Economic Literature: Analysis of an Auction Marketplace," *Review of Industrial Organization*, 37(1), 3–42.
- Hernando-Veciana, A., and F. Michelucci (2011): "Second Best Efficiency and the English Auction," *Games and Economic Behavior*, 73(2), 496–506.
- Hernando-Veciana, Á., and F. Michelucci (2013): "Do Not Panic: How to Avoid Inefficient Rushes Using Multi-stage Auctions," *CERGE-EI Working Paper Series* 489, CERGE-EI.
- Houser, D., and J. Woolders (2006): "Reputation in Auctions: Theory, and Evidence from eBay," *Journal of Economics & Management Strategy*, 15(2), 353–369.
- Jehiel, P., and L. Lamy (2015a): "On Absolute Auctions and Secret Reserve Prices," *The RAND Journal of Economics*, 46(2), 241–270.
- Jehiel, P., and L. Lamy (2015b): "On Discrimination in Auctions with Endogenous Entry," *The American Economic Review*, 105(8), 2595–2643.
- Jeziorski, P., and I. Segal (2015): "What Makes Them Click: Empirical Analysis of Consumer Demand for Search Advertising," *American Economic Journal: Microeconomics*, 7(3), 24–53.
- Jin, G.Z., and A. Kato (2006): "Price, Quality, and Reputation: Evidence from an Online Field Experiment," *The RAND Journal of Economics*, 37(4), 983–1005.
- Kauffman, R.J., and C.A. Wood (2005): "The Effects of Shilling on Final Bid Prices in Online Auctions," *Electronic Commerce Research and Applications*, 4(1), 21–34.
- Klemperer, P. (2004): *Auctions: Theory and Practice*, The Toulouse Lectures in Economics. Princeton, NJ: Princeton University Press.
- Krishna, V. (2002): *Auction Theory*. 1st edition. Cambridge, MA: Academic Press.
- Lamy, L. (2009): "The Shill Bidding Effect versus the Linkage Principle," *Journal of Economic Theory*, 144(1), 390–413.
- Levin, D., and J.L. Smith (1994): "Equilibrium in Auctions with Entry," *The American Economic Review*, 84(3), 585–599.
- Liu, D., and J. Chen (2006): "Designing Online Auctions with Past Performance Information," *Decision Support Systems*, 42(3), 1307–1320.
- Liu, D., J. Chen, and A.B. Whinston (2010): "Ex Ante Information and the Design of Keyword Auctions," *Information Systems Research*, 21(1), 133–153.
- Livingston, J.A. (2005): "How Valuable is a Good Reputation? A Sample Selection Model of Internet Auctions," *Review of Economics and Statistics*, 87(3), 453–465.
- Lucking-Reiley, D., D. Bryan, N. Prasad, and D. Reeves (2007): "Pennies from eBay: The Determinants of Price in Online Auctions," *The Journal of Industrial Economics*, 55(2), 223–233.
- Malakhov, A., and R.V. Vohra (2008): "Optimal Auctions for Asymmetrically Budget Constrained Bidders," *Review of Economic Design*, 12(4), 245–257.
- Malmendier, U., and Y.H. Lee (2011): "The Bidder's Curse," *The American Economic Review*, 101(2), 749–787.
- Maskin, E. (1992): "Auctions and Privatization," in H. Siebert (ed.), *Privatization: Symposium in Honour of Herbert Giersch*, Institute für Weltwirtschaft an der Universität Kiel.
- Maskin, E.S. (2000): "Auctions, Development, and Privatization: Efficient Auctions with Liquidity-constrained Buyers," *European Economic Review*, 44, 667–681.
- Maskin, E., and J. Riley (2000): "Asymmetric Auctions," *The Review of Economic Studies*, 67(3), 413–438.
- Mathews, T., and B. Katzman (2006): "The Role of Varying Risk Attitudes in an Auction with a Buyout Option," *Economic Theory*, 27(3), 597–613.
- McAfee, P. (1993): "Mechanism Design by Competing Sellers," *Econometrica*, 61(6), 1281–1312.
- McDonald, C.G., and V.C. Slawson (2002): "Reputation in an Internet Auction Market," *Economic Inquiry*, 40(4), 633–650.
- Melnik, M.I., and J. Alm (2002): "Does a Seller's Ecommerce Reputation Matter? Evidence from eBay Auctions," *The Journal of Industrial Economics*, 50(3), 337–349.
- Menezes, F.M., and P.K. Monteiro (2008): *An Introduction to Auction Theory*. New York: Oxford University Press.

- Milgrom, P.R. (1985): "The Economics of Competitive Bidding: A Selective Survey," in L. Hurwicz, D. Schmeidler, and H. Sonnenschein (eds), *Social Goals and Social Organization: Essays in Memory of Elisha Pazner*, Cambridge, UK: Cambridge University Press, pp. 261–292.
- Milgrom, P. (2000): "Putting Auction Theory to Work: The Simultaneous Ascending Auction," *Journal of Political Economy*, 108(2), 245–272.
- Milgrom, P. (2004): *Putting Auction Theory to Work*. 1st edition. Cambridge, UK: Cambridge University Press.
- Milgrom, P., and I. Segal (2002): "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70(2), 583–601.
- Milgrom, P., and R. Weber (1982): "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50, 1089–1122.
- Moreno, D., and J. Wooders (2011): "Auctions with Heterogeneous Entry Costs," *The RAND Journal of Economics*, 42(2), 313–336.
- Myerson, R.B. (1981): "Optimal Auction Design," *Mathematics of Operation Research*, 6(1), 58–73.
- Ockenfels, A., and A.E. Roth (2006): "Late and Multiple Bidding in Second Price Internet Auctions: Theory and Evidence Concerning Different Rules for Ending an Auction," *Games and Economic Behavior*, 55(2), 297–320.
- Peters, M., and S. Severinov (2006): "Internet Auctions with Many Traders," *Journal of Economic Theory*, 130(1), 220–245.
- Pitchik, C. (2009): "Budget-constrained Sequential Auctions with Incomplete Information," *Games and Economic Behavior*, 66, 928–949.
- Pitchik, C., and A. Schotter (1988): "Perfect Equilibria in Budget-constrained Sequential Auctions: An Experimental Study," *RAND Journal of Economics*, 19(3), 363–389.
- Resnick, P., and R. Zeckhauser (2002): "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System," *The Economics of the Internet and E-commerce*, 11(2), 23–25.
- Reynolds, S.S., and J. Wooders (2009): "Auctions with a Buy Price," *Economic Theory*, 38, 9–39.
- Rhodes-Kropf, M., and S. Viswanathan (2005): "Financing Auction Bids," *RAND Journal of Economics*, 36(4), 789–815.
- Roth, A.E., and A. Ockenfels (2002): "Last Minute Bidding and the Rules for Ending Second Price Auctions: Evidence from eBay and Amazon Auctions on the Internet," *American Economic Review*, 92(4), 1093–1103.
- Schneider, H.S. (2015): "The Bidder's Curse: Comment," *American Economic Review*, 106(4), 1182–1194.
- Shahriar, Q., and J. Wooders (2011): "An Experimental Study of Auctions with a Buy Price under Private and Common Values," *Games and Economic Behavior*, 72(2), 558–573.
- Varian, H.R. (2007): "Position Auctions," *International Journal of Industrial Organization*, 25(6), 1163–1178.
- Vickrey, W. (1961): "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, 16, 8–37.
- Vincent, D.R. (1995): "Bidding Off the Wall: Why Reserve Prices May Be Kept Secret," *Journal of Economic Theory*, 65(2), 575–584.
- Wilcox, R.T. (2000): "Experts and Amateurs: The Role of Experience in Internet Auctions," *Marketing Letters*, 11(4), 363–374.
- Wilson, R. (1987): "Game-theoretic Approaches to Trading Processes," in T. Fassett Bewley (ed.), *Advances in Economic Theory: Fifth World Congress of the Econometric Society*, Cambridge, UK: Cambridge University Press.
- Yang, S., and A. Ghose (2010): "Analyzing the Relationship between Organic and Sponsored Search Advertising: Positive, Negative, or Zero Interdependence?" *Marketing Science*, 29(4), 602–623.
- Yeon-Koo Che, J.K. (2010): "Bidding with Securities: Comment," *The American Economic Review*, 100(4), 1929–1935.
- Zheng, C. (2001): "High Bids and Broke Winners," *Journal of Economic Theory*, 100(1), 129–171.



13. Differential oligopoly games in environmental and resource economics

Luca Lambertini

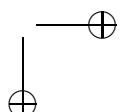
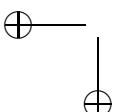
1 INTRODUCTION

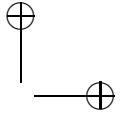
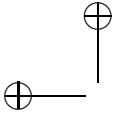
The economic theory of natural resources and the environment is characterized by an intensive use of dynamic analysis for intuitive reasons, as the time dimension is essential to our understanding of the impact of complex economic systems on the preservation of resources and species as well as on the quality of the environment and climate change.

In this respect, the theory of oligopoly makes no exception. Indeed, the territory located at the interception between industrial organization and environmental and resource economics is an area in which the application of dynamic game theory has been very fertile over the last few decades. The large body of available research in this direction convincingly illustrates what dynamic game theory can do to refine our understanding of such a complex and crucial matter, in terms of positive as well as normative analysis. The aim of this chapter is to offer a comprehensive overview of the resulting literature based on differential games (that is, dynamic games in continuous time), whose general focus is on the interplay between either regulated or unregulated oligopolistic firms' profit incentives and the preservation of the stock of natural capital.¹

The remainder is structured as follows. Section 2 sets the stage for Cournot oligopoly games with either polluting emissions or resource extraction under open-loop rules, without any form of regulation on emissions or access to the commons. Section 3 reviews the literature on environmental games with feedback structures, where firms may be subject to emission taxes and possibly activate R&D projects for green technologies. Games involving the exploitation of renewable and non-renewable resources are examined in Section 4. In the related literature, a considerable amount of attention has been devoted to the determination of the optimal number of firms in the commons, cartel behaviour, exploration and the introduction of alternative technologies. Section 5 is devoted to the scant literature analysing the interplay between emissions and natural resources in a single framework. The few existing results on two recent offshoots, namely, corporate environmentalism and the Porter hypothesis, are reported in Section 6. The bearing of international trade on the environment – itself a crucial issue in the ongoing debate on globalization and climate change – is reviewed in Section 7.

¹ For additional surveys on the same topics, see Jørgensen, Martín-Herrán and G. Zaccour (2010) Long (2010, 2011) and Lambertini (2013).





2 PRELIMINARIES

A few essential features of the nature of differential oligopoly games either with polluting emissions or with the exploitation of a natural resource can be grasped as follows.

The model unravels over continuous time $t \in [0, \infty]$. The market is supplied by $n \geq 2$ firms producing a homogeneous good, whose inverse demand function is $p(t) = a - Q(t)$ at any time t , where $a > 0$ is the time-invariant reservation price of the representative consumer and $Q(t) = \sum_{i=1}^n q_i(t)$ is the sum of individual outputs $q_i(t)$. Firms share the same technology, characterized by marginal cost $c \in (0, a)$, constant over time. Firms operate without any fixed costs.

If the focus is on polluting emissions (generated – say – by production), what matters is the evolution of the stock of emissions $S(t)$ through to the following dynamics:

$$\dot{S}(t) = Q(t) - \delta S(t), \quad (13.1)$$

where $\delta > 0$ is the constant decay rate characterizing the natural carbon sinks and, for the sake of simplicity, the rate of CO₂-equivalent emissions per unit of output is normalized to one, so that the instantaneous amount of pollutants per firm is $s_i(t) = q_i(t)$.

Consumers have no environmental concerns and there exists no environmental regulation. As a consequence, firms have no reason to undertake costly R&D projects for cleaner technologies. The instantaneous profit function of firm i is $\pi_i(t) = [p(t) - c] q_i(t)$, and each firm i chooses $q_i(t)$ non-cooperatively, so as to maximize the discounted profit flow

$$\Pi_i = \int_0^\infty [p(t) - c] q_i(t) e^{-\rho t} dt, \quad (13.2)$$

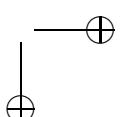
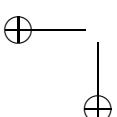
under the constraints posed by the state equation (13.1) and the initial condition $S(0) = S_0 \geq 0$. Parameter $\rho > 0$ represents a constant discount rate common to all firms and the policy maker, who evaluates the performance of this industry on the basis of the instantaneous social welfare function is

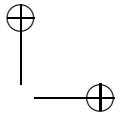
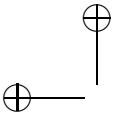
$$SW(t) = \sum_{i=1}^n \pi_i(t) + CS(t) - D(S(t)) \quad (13.3)$$

where $CS(t) = Q^2(t)/2$ is consumer surplus and aggregate emissions $S(t)$ cause the environmental damage $D(S(t))$, which is commonly assumed to be a quadratic function of both $S(t)$. This is the simplest representation of the damage, which could include an additional effect related to industry output, and in fact it is often modelled to include it (see below).

In the alternative scenario, firms jointly exploit a (possibly renewable) natural resource (say, a common pool), whose dynamic behaviour is governed by the following state equation:

$$\dot{X} = F(X(t)) - Q(t) \quad (13.4)$$





where

$$F(X(t)) = \begin{cases} \eta X(t) & \forall X(t) \in [0, X_y] \\ \eta X_y \left(\frac{X_{\max} - X(t)}{X_{\max} - X_y} \right) & \forall X(t) \in [X_y, X_{\max}] \end{cases} \quad (13.5)$$

In (13.5), $X(t)$ is the resource stock, $\eta \geq 0$ is its *implicit* growth rate when the stock is at most equal to X_y and ηX_y is the maximum sustainable yield. (13.4–13.5) jointly entail that if the resource stock is small enough, the population grows at an exponential rate; and beyond X_y , the asset grows at a decreasing rate. Additionally, X_{\max} identifies the *carrying capacity* of the habitat: above this threshold, the rate of growth of the resource becomes negative, since it is limited by the available amounts of food and space. In the remainder, we will confine our attention to the case in which $F(X(t)) = \eta X(t)$.² Clearly, if $\eta = 0$ the resource is non-renewable (as is the case for fossil fuels).

Again, if the representative consumer does not explicitly care for resource or species preservation, demand is defined as above. If, in addition, access to the commons is unregulated, then each firm takes $\pi_i(t) = [p(t) - c] q_i(t)$ as its instantaneous objective function, and chooses $q_i(t)$ to maximize (13.2) in the non-cooperative game over time $t \in [0, \infty]$.

As for the government or its agency, the relevant instantaneous welfare function is written as follows:

$$SW(t) = \sum_{i=1}^n \pi_i(t) + CS(t) + X(t) \quad (13.6)$$

where the presence of $X(t)$ testifies in favour of the public authority's concern about the preservation of the natural resource.

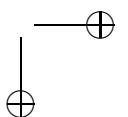
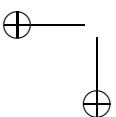
Taking $F(X(t)) = \eta X(t)$, both scenarios define linear state games in which open-loop rules deliver subgame perfect (more precisely, strongly time consistent or feedback) controls.³ This property can be easily grasped by examining the solution of the open-loop problem of firm i in the two settings.

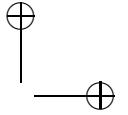
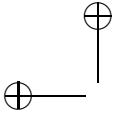
When the relevant state equation is (13.1), the representative firm's Hamiltonian function is

$$\begin{aligned} \mathcal{H}_i(t) = & e^{-\rho t} \left\{ \left(a - q_i(t) - \sum_{j \neq i} q_j(t) - c \right) q_i(t) \right. \\ & \left. + \lambda_i(t) \left[q_i(t) + \sum_{j \neq i} q_j(t) - \delta S(t) \right] \right\} \end{aligned} \quad (13.7)$$

² This amounts to saying that the dominant approach to the problem of resource extraction in dynamic oligopoly models uses a linearized version of the original Lotka-Volterra formulation (Lotka, 1925; Volterra, 1931) of the resource dynamics.

³ For more on this as well as other classes of games where the open-loop solution is a degenerate feedback one, see Dockner, Feichtinger and Jørgenson (1985), Fershtman (1987), Mehlmann (1988, ch. 4), Dockner et al. (2000, ch. 7) and Cellini, Lambertini and Leitmann (2005), *inter alia*.





in which $\lambda_i(t) = e^{\rho t} \gamma_i(t)$ is the co-state variable in current value associated with the dynamics of polluting emissions. From (13.7), one derives the following set of necessary conditions:⁴

$$\frac{\partial \mathcal{H}_i}{\partial q_i} = e^{-\rho t} \left(a - c - 2q_i - \sum_{j \neq i} q_j + \lambda_i \right) = 0 \quad (13.8)$$

$$\dot{\lambda}_i = -\frac{\partial \mathcal{H}_i}{\partial S} + \rho \lambda_i \Leftrightarrow \dot{\lambda}_i = (\delta + \rho) \lambda_i \quad (13.9)$$

Given that (13.9) obviously admits the solution $\lambda_i = 0$ at any instant t , this immediately entails that the quasi-static Cournot equilibrium with $q^{OL} = q^{CN} = (a - c) / (n + 1)$ emerges at all times. The required transversality condition $\lim_{t \rightarrow \infty} e^{-\rho t} \lambda_i = 0$ is also met. The resulting level of pollution at the steady state is $S^{OL} = nq^{OL}/\delta = n(a - c) / [(n + 1)\delta]$, which is monotonically increasing and concave in the number of firms.

On the basis of the dynamic properties of the state variable S , one can then verify that the steady state equilibrium (S^{OL}, q^{OL}) reached under open-loop information is a saddle point. This is shown graphically in Figure 13.1, where the arrows appearing along the horizontal line corresponding to q^{OL} illustrate the convergence towards the intersection with the locus $\dot{S} = 0$, along which $S = nq/\delta$.

Relying on the saddle point stability property, a public authority evaluating the welfare performance (or specifically the level of the environmental damage) stemming from firms' behaviour may confidently rely on the fact that this industry – if unregulated – will indeed follow a stable path towards $S^{OL} = nq^{OL}/\delta$ and use the welfare function (13.3), for instance, to shape its environmental policy adequately.

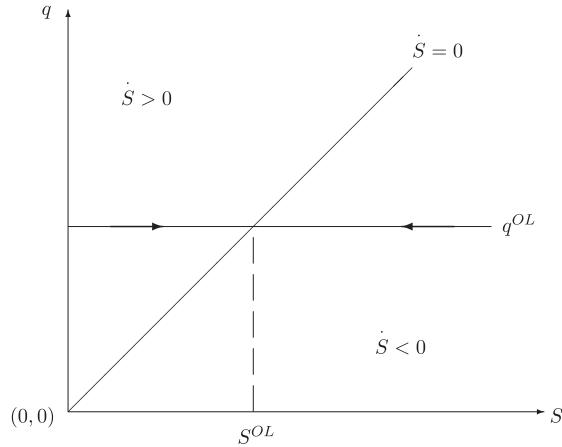
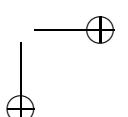
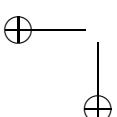


Figure 13.1 Open-loop solution: phase diagram in the (S, q) space

⁴ In the remainder of the exercise, the explicit indication of the time argument is omitted for the sake of brevity.





We may now turn our attention to the model describing the extraction of the natural resource. In this case, the relevant state equation is (13.4), and firm i must choose $q_i(t)$ non-cooperatively to maximize the Hamiltonian function:

$$\begin{aligned} \mathcal{H}_i(t) = & e^{-\rho t} \left\{ \left(a - q_i(t) - \sum_{j \neq i} q_j(t) - c \right) q_i(t) \right. \\ & \left. + \lambda_i(t) \left[\eta X(t) - q_i(t) - \sum_{j \neq i} q_j(t) \right] \right\} \end{aligned} \quad (13.10)$$

The necessary conditions deriving from (13.10) are:

$$\frac{\partial \mathcal{H}_i}{\partial q_i} = e^{-\rho t} \left(a - c - 2q_i - \sum_{j \neq i} q_j - \lambda_i \right) = 0 \quad (13.11)$$

$$\dot{\lambda}_i = -\frac{\partial \mathcal{H}_i}{\partial S} + \rho \lambda_i \Leftrightarrow \dot{\lambda}_i = (\rho - \eta) \lambda_i \quad (13.12)$$

As above, the open-loop solution is strongly time consistent because the co-state equation (13.12) is a differential equation in separable variables admitting the nil solution at all times. As a consequence, the same quasi-static Cournot-Nash equilibrium is adopted at every instant by all firms alike. Given $q^{OL} = (a - c) / (n + 1)$, we might expect to observe the dynamic system reaching the steady state at $X^{OL} = nq^{OL}/\eta$, as depicted in Figure 13.2.

Unfortunately, we should be prepared to observe completely different outcomes, since the state dynamics reveals that the point (X^{OL}, q^{OL}) is unstable, as implied by the arrows along the horizontal line at q^{OL} , reflecting the sign of \dot{X} above and below the locus $X = nq/\eta$. More

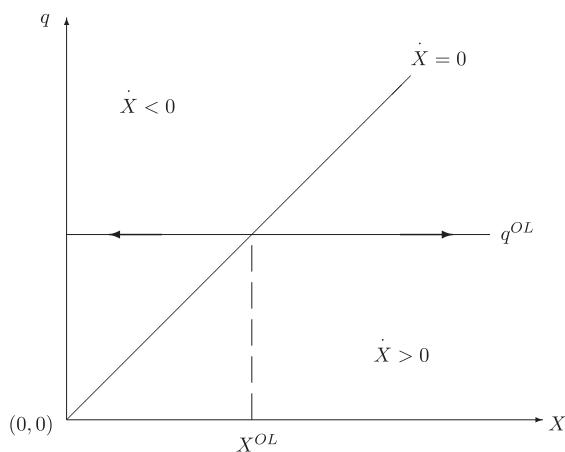
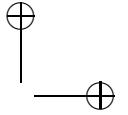
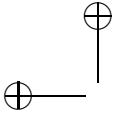


Figure 13.2 Open-loop solution: phase diagram in the (X, q) space





precisely, the outcome will depend on the initial stock of the resource: if $X_0 \in (0, X^{OL})$, the resource stock (or population) will shrink to zero; if instead $X_0 > X^{OL}$, the stock will grow indefinitely. Consequently, a public agency in charge of regulating this industry to ensure the preservation of the natural resource cannot tailor its policy instrument on the presumption that the industry will be converging to (X^{OL}, q^{OL}) . Likewise, firms cannot rely on a permanent rent granted by resource extraction and attained through open-loop rules, if the initial resource stock is below X^{OL} .

This simple exercise, based on the replication of the Cournot-Nash behaviour by firms over time, reveals that there exists a striking difference between a Cournot oligopoly producing an environmental externality and one exploiting a natural resource. Although both models seemingly share the same essential structure (a quadratic profit function coupled with a linear state dynamics, and open-loop information), the qualitative and quantitative features of the final outcome are opposite, in such a way that, when the survival of a natural resource is at stake, firms and regulators are equally aware that the fate of the resource and that of the industry also ultimately depend on the initial condition, as open-loop rules cannot ensure stability. The emergence of saddle point stability and the lack thereof, respectively, have largely affected the way the two problems have been treated in more comprehensive models encompassing the optimal design of policy instruments and the adoption of feedback information.

3 POLLUTING EMISSIONS

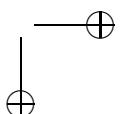
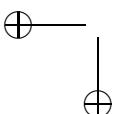
The cornerstone of the debate on environmental externalities in differential oligopoly games is the model by Benchekroun and Long (1998), which encompasses previous contributions focusing solely on monopoly (Bergstrom, Cross and Porter, 1981; and Karp and Livernois, 1994).⁵

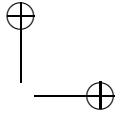
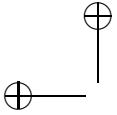
Benchekroun and Long (1998) admit the presence of a regulator introducing an emission tax affecting the optimal production plans of firms. By doing so, they prove the existence of a unique tax rate imposed on polluting emission that drives oligopolistic firms (i) to replicate the socially efficient path and therefore (ii) to reach exactly the same steady state that would be attained under social planning. In Benchekroun and Long (1998) this is proved in a general setting ensuring the concavity of the model, while for expositional purposes I am using the same setup as in the previous section (which in the original paper is used as an illustrative example), with $D(S(t)) = \gamma S^2(t)/2$, where γ is a positive constant.

The first step consists in describing the command optimum in which the public authority chooses outputs to maximize the discounted welfare flow generated by function (13.3). Hence, the relevant Hamiltonian function is

$$\mathcal{H}(t) = e^{-\rho t} [\pi_i(t) + CS(t) - D(S(t)) + \lambda \dot{S}] \quad (13.13)$$

⁵ In a later note, Benchekroun and Long (2002a) revisit the monopoly setting to prove the existence of a continuum of linear Markovian tax rules inducing the firm to achieve the socially efficient path.





Manipulating the necessary conditions, one obtains:

$$Q^{SP} = \frac{(a - c)(\delta + \rho) - \gamma S}{\delta + \rho} \quad (13.14)$$

$$\lambda^{SP} = -a + c + Q^{SP} \quad (13.15)$$

where superscript *SP* stands for *social planning*. The above expressions reveal that the socially optimal individual (and therefore also aggregate) output is decreasing in the amount of aggregate emission at any time, and the shadow price attribute to the state variable is clearly negative, as one would expect in the presence of a negative externality.⁶ The resulting steady state level of emissions is

$$S^{SP} = \frac{(a - c)(\delta + \rho)}{\delta(\delta + \rho) + \gamma} \quad (13.16)$$

Concerning S^{SP} , it is worth noting that it is positive, increasing in market size $a - c$, and decreasing in γ . In particular, the latter property cannot be expected to apply if profit-maximizing firms are free to choose output levels in absence of any regulation, as they will not internalize the external effect if not compelled to do so.

Suppose the policy maker introduces a linear tax rule $\tau(S) = \alpha S + \beta$, in such a way that the instantaneous cost function of firm i rewrites as $C_i = [c + \tau(S)] q_i$. Firms follow open-loop rules but are subject to an environmental policy increasing their marginal and average production cost. The Hamiltonian of the representative firm is therefore the following:

$$\begin{aligned} \mathcal{H}_i(t) = & e^{-\rho t} \left\{ \left(a - q_i(t) - \sum_{j \neq i} q_j(t) - c - \tau(S) \right) q_i(t) \right. \\ & \left. + \lambda_i(t) \left[q_i(t) + \sum_{j \neq i} q_j(t) - \delta S(t) \right] \right\}, \end{aligned} \quad (13.17)$$

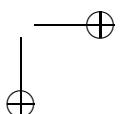
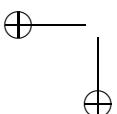
which yields the set of necessary conditions

$$\frac{\partial \mathcal{H}_i}{\partial q_i} = e^{-\rho t} \left(a - c - \alpha S - \beta - 2q_i - \sum_{j \neq i} q_j + \lambda_i \right) = 0 \quad (13.18)$$

$$\dot{\lambda}_i = -\frac{\partial \mathcal{H}_i}{\partial S} + \rho \lambda_i \Leftrightarrow \dot{\lambda}_i = \alpha q_i + (\delta + \rho) \lambda_i \quad (13.19)$$

Evidently, (13.19) does not admit the nil solution with respect to (w.r.t.) λ_i any more. Hence, one has to proceed as follows. Imposing symmetry across controls, and solving (13.18) w.r.t.

⁶ To verify that $\lambda^{SP} < 0$, one may simply observe that Q^{SP} , accounting for polluting emissions, is lower than the perfectly competitive per firm output $(a - c)/n$.



to q , one obtains the expression of the optimal individual output as a function of the co-state variable λ and the tax rule (itself defined in terms of the stock of industry emissions) at any time t , $q^{CN}(\lambda, \tau(S))$. Plugging this expression into (13.19) yields the co-state equation as a differential equation in λ . Now, in order for the population of firms to replicate the same path and reach the same outcome as the social planner, their collective intertemporal production plans and the associated shadow price must coincide with those the benevolent planner would choose if he had to be in their shoes. Imposing these two conditions delivers a unique pair $(\alpha^{OL}, \beta^{OL})$ shaping the efficient tax rule $\tau(S)$ in the case of open-loop strategies (Benckroun and Long, 1998, Proposition 2, p. 334). An analogous conclusion holds if firms adopt feedback strategies of the type $q_i = f(S)$ (Benckroun and Long, 1998, Proposition 6, p. 338), and we may denote the resulting pair as (α^F, β^F) , where superscript F stands for *feedback*. However, the uniqueness of the optimal tax policy should not be taken with too much confidence, since, in order to adopt the correct regulatory instrument the authority should in the first place ascertain the information structure on the basis of which firms will design their output strategies.

If this is too demanding a requirement, the regulator may opt for a more direct alternative consisting of nationalizing at least one of the firms operating in this industry and use these public firms – instead of an emission tax or other tools such as pollution quotas or environmental standards – as endogenous instruments to regulate the behaviour of the entire sector. This is the subject matter of Dragone, Lambertini and Palestini (2014) using the same layout as in Benckroun and Long (1998) modified in a single respect by assuming that $m \in [1, n - 1]$ firms be publicly held and therefore adopt welfare-maximizing production plans. This detail, *per se*, is sufficient to ensure that private (profit-seeking) firms will be forced to internalize the externality, as the latter appears explicitly in every public firm's objective and first-order conditions. In order to make the model more realistic, Dragone et al. (2014) allow for the presence of X-inefficiency in public units (Leibenstein, 1966), whereby private firms enjoy a marginal cost advantage. In this setting, it clearly emerges that a public firm must strike a non-trivial balance between consumer surplus on one side and the need to affect private firms' behaviour (and therefore the external effect) on the other, as the latter is the explicit reason for the nationalization of some of the productive units in the industry. Under feedback rules, the output of public firms as well as the industry output are monotonically decreasing in the level of aggregate emissions. Both along the equilibrium path and at the steady state equilibrium, the mixed oligopoly produces more than a social planner would do, except when X-inefficiency is altogether absent. In such a case, nationalizing a single firm suffices to replicate the first best outcome.

Another extension of the same model tackles an issue we shall encounter again below, discussing the literature concerning the exploitation of natural resources, namely, cartel behaviour. Intuition suggests that, since a cartel, irrespective of whether it is based on implicit collusion or explicit cooperation, usually restricts industry output as compared to the non-cooperative Cournot-Nash equilibrium, this should bring about a reduction in polluting emissions as these are monotonic in industry output. Admitting the emergence of a cartel involving all firms in the industry, Benckroun and Chaudhury (2011) show that a tax rule linear in aggregate emissions indeed stabilizes the cartel in the sense of d'Aspremont et al. (1983), Donsimoni (1985) and Donsimoni, Economides and Polemarchakis (1986). Their analysis also proves that, while the tax induces a decrease in the stock of pollutants, its contribution to cartel stability translates into a welfare decrease as shrinking output thwarts consumer surplus.

To conclude this overview, it is worth stressing that most of the literature discussed here as well as in the remainder of the survey, where natural resources are considered, relies on models in which productive capacity and its build-up process are left out of the picture – surprisingly enough, as there is an obvious interplay between the extant firms' capacity endowment and the environmental impact of production. An exception is the Cournot-Ramsey model in Dragone, Lambertini and Palestini (2010), where it is shown that, if polluting emissions are generated by production – and therefore depend on installed capacity – then the Ramsey golden rule (Ramsey, 1928) may no longer be socially optimal and the state-control system may indeed become explosive.

3.1 Green R&D

While all of the aforementioned contributions model the relationship between firms' output strategies and the dynamics of the external effect, a few recent ones include R&D efforts for clean(er) technologies.⁷ At odds with the conventional view holding that firms should not be expected to invest in green R&D unless spurred on by regulation (such as the emission tax considered in Benckroun and Long, 1998), Dragone, Lambertini and Palestini (2013) use the same baseline dynamic Cournot setup to find that unregulated firms might well perform such activities even in absence of a tax on emissions, via an indirect effect generated by the fact that R&D efforts may indeed be appealing to firms if such efforts go along with an output contraction closely resembling cartel behaviour, although the setup remains fully non-cooperative.

A completely different aspect, closely connected with the whole literature concerning the relationship between the intensity of market competition and the shape of aggregate R&D stemming from the early Schumpeter-Arrow (Schumpeter, 1942; Arrow, 1962) debate, is the core of the analysis in Feichtinger et al. (2016), where emissions are taxed but market price is a regulated constant, in the hands of a regulator. The equilibrium of the resulting differential game reproduces a result that has recently been highlighted in growth theory, namely, the emergence of an inverted-U aggregate R&D expenditure at equilibrium (Aghion et al., 2005). In a model where R&D is environmentally targeted, this result implies the existence of a unique industry structure maximizing the greenness of the production technology in use, pointing out the need for coordinating environmental policies and the regulation of the entry process in virtually any sector with a non-negligible environmental impact.

4 NATURAL RESOURCE EXTRACTION

The nature of the long-standing discussion on natural resource exploitation is twofold, as it depends on whether the resource at stake is renewable or non-renewable, although some of its essential features are invariant with respect to this characteristic. Intuitively, as we shall see in the remainder of this section, the stream of research on non-renewables has repeatedly addressed the implications of cartel behaviour.

⁷ The issue of green R&D incentives and the design of the appropriate regulatory instruments to enhance them has been extensively investigated in static (multistage) games. For surveys, see Requate (2005) and Lambertini (2013, ch. 2).

4.1 Renewable Resources

The differential game approach to renewables has a long tradition, which I will try to reconstruct here by offering first a brief summary of the discussion about the so-called fish wars, switching thereafter to the parallel and still very lively literature on the oligopolistic exploitation of common pool resources.

4.1.1 The early debate about fishery games

This stream of research is in very close connection with the tragedy of the commons (Gordon, 1954; Hardin, 1968)⁸ and the Verhulst-Lotka-Volterra model (Verhulst, 1838; Lotka, 1925; Volterra, 1931). The basic model dates back to Clark and Munro (1975) and Levhari and Mirman (1980), and has the unpleasant drawback of relying mostly on open-loop rules, as characterizing feedback strategies is very difficult in view of the population dynamics à la Lotka-Volterra:

$$\dot{X} = \alpha X(t) [1 - \beta X(t)] - Q(t), \quad (13.20)$$

over $t \in [0, \infty]$. Parameters α and β are both positive, and $Q(t)$ is the instantaneous harvest by the n fishermen sharing the same fishing ground. As (13.20) is quadratic in the state (while being linear in all controls), the model does not identify a linear-quadratic differential game. Under open-loop information, the tragedy emerges in the form of each player's lack of internalization of the fact that his harvest today will be reflected in the rivals' behaviour via the change in the stock.⁹ The most notable exception, where feedback rules are considered, is the work of Levhari and Mirman (1980), where it is shown that the tragedy also emerges if the harvesting function is independent of the stock.¹⁰ The bulk of the current debate is based on a linear version of dynamics (13.20).

4.1.2 Oligopoly and the commons

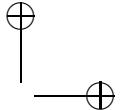
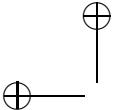
Here I will summarize the main findings of a rich debate about the impact of oligopolistic firms on the survival of a renewable resource (or species), on the basis of a compound of several contributions by Benckroun (2003, 2008), Fujiwara (2008), Lambertini and Mantovani (2014, 2016), Colombo and Labrecciosa (2015) and Lambertini (2016). All of this stream of research can be put together using the benchmark model based on the Cournot model with n firms selling a homogeneous good whose production relies on the extraction of a resource following the state equation (13.4), with its growth rate obeying (13.5).

Access to the common resource pool being unregulated, the open-loop equilibrium is as in Section 2: note that, as it replicates the static Cournot-Nash behaviour, the resulting individual and aggregate extraction rates are obviously independent of both the discount rate ρ and

⁸ See also Clark (1973, 1990).

⁹ That is, the tragedy arises because the harvesting strategy is a function of the fish stock and players don't take this feature into account. If the strategy does not depend on the stock, then the open-loop solution is a degenerate feedback one and therefore Markov perfect (see Chiarella et al., 1984).

¹⁰ The literature referring to the fishery game is quite extensive: other relevant contributions include Clemhout and Wan (1985a), Fischer and Mirman (1986), Cave (1987), Benhabib and Radner (1992), Chichilnisky (1994), Dockner and Sorger (1996), Brander and Taylor (1997), Dawid and Kopel (1997) and Benckroun and Long (2002b), among many others.



the resource growth rate η . For the moment, assume $\eta > \rho(n^2 + 1)/2$, which ensures the positivity of the residual resource stock at the steady state under linear feedback rules.

Under feedback information, firm i chooses its extraction rate q_i to maximize its Hamilton-Jacobi-Bellman equation:

$$\rho V_i(X) = \max_{q_i} [(a - c - Q) q_i + V'_i(X) (\eta X - Q)] \quad (13.21)$$

where $V_i(X)$ is firm i 's value function and $V'_i(X) = \partial V_i(X) / \partial X$ is its partial derivative w.r.t. the state variable X . Firms play simultaneously and non-cooperatively at all times $t \in [0, \infty]$. The first-order condition (FOC) taken w.r.t. the individual control variable q_i is

$$a - c - 2q_i - \sum_{j \neq i} q_j - V'_i(X) = 0 \quad (13.22)$$

In view of the *ex ante* symmetry across firms, one can impose the condition $q_j = q_i = q$ and solve (13.22) to obtain $q = [a - c - V'(X)] / (n + 1)$. If $a - c > V'(X)$, one can plug this expression into (13.21), guess a linear-quadratic form for the value function, $V(X) = \varepsilon_1 X^2 + \varepsilon_2 X + \varepsilon_3$, where coefficients $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are to be determined, and then solve (13.21) via the associated system of three Riccati equations. This delivers two steady state equilibria, one replicating the above open-loop solution (thereby showing it is indeed a degenerate feedback equilibrium), while the other is attained at

$$X^{LF} = \frac{nq^{LF}}{\eta} = \frac{(a - c)[2\eta - \rho(n^2 + 1)]}{\delta[2\eta - \rho(n + 1)](n + 1)} > 0 \quad \forall \eta > \frac{\rho(n^2 + 1)}{2} \quad (13.23)$$

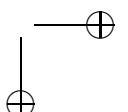
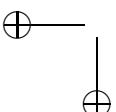
being generated by a proper linear feedback rule holding at any time t :

$$q^{LF}(X) = \frac{\eta(2\eta - \rho)(n + 1)^2 X - (a - c)[2\eta - \rho(n^2 + 1)]}{2\eta(n + 1)n^2} \quad (13.24)$$

so that the aggregate extraction is $Q^{LF}(X) = nq^{LF}(X)$. Both extraction rules and the resulting steady state levels of the resource stock are represented in Figure 13.3, where the arrows along q^{OL} and $q^{LF}(X)$ illustrate the stability properties of each strategy. In particular, it appears that the linear feedback solution is indeed stable.

The linear feedback strategy is not the unique feedback solution. In fact, there exist a continuum of non-linear feedback equilibria, which can be generated in two alternative and not equivalent ways. The first to appear in the literature (Tsutsui and Mino, 1990) rests on restrictive conditions limiting the analysis to a bounded region of the state variable (therefore imposing boundaries on payoff functions and control variables), while the second, introduced by Rowat (2007) on the basis of the "catching up optimality" criterion (Dockner et al., 2000, pp. 62–7) is a general one and spans the entire state space.

Without delving into the technical details of the matter, here I will confine myself to a synthetic exposition of the derivation of the infinitely many non-linear solutions of the Cournot extraction game. The first step consists in going back to FOC (13.22) and solve it w.r.t. the slope of the value function, to obtain $V'(X) = a - c - (n + 1)q(X)$. Again,



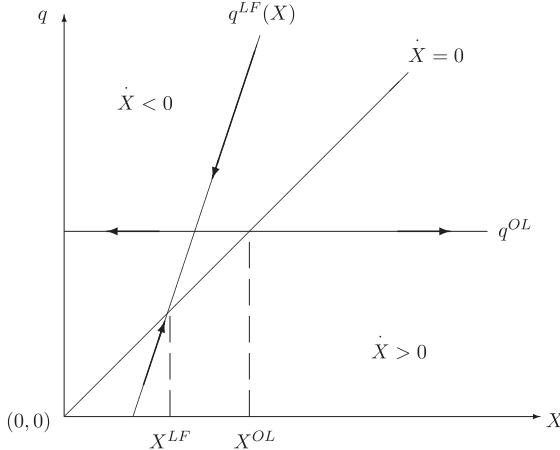


Figure 13.3 Open-loop and linear feedback solutions in the (X, q) space

this is admissible provided that $a - c > V'(X)$. Substituting this expression into (13.21), differentiating both sides of it with respect to X and rearranging terms, one may write the following differential equation:

$$q'(X) = \frac{(\eta - \rho)[(n + 1)q(X) - (a - c)]}{2n^2q(X) - \eta(n + 1)X - (n - 1)(a - c)}, \quad (13.25)$$

which implicitly identifies any feedback strategy, including the open-loop one and $q^{LF}(X)$ (see Fujiwara, 2008, p. 218; Lambertini and Mantovani, 2014, pp. 117–18).

Then, using the notion that, at the tangency point between the generic firm's highest isocline and the state stability locus it must be $q'(S) = \delta/n$, one identifies the degenerate non-linear feedback equilibrium at

$$X^T = \frac{(a - c)(\eta - n\rho)}{\eta[2\eta - \rho(n + 1)]} = \frac{nq^T}{\eta} \quad (13.26)$$

where superscript T stands for tangency. The tangency solution under non-linear feedback rules is point \mathbb{T} in Figure 13.4 (which appears in Fujiwara 2008, Lambertini, 2016; and Lambertini and Mantovani, 2016), in which the steady states generated by open-loop and linear feedback rules are identified as \mathbb{OL} and \mathbb{LF} , respectively. In the same figure, it also appears that an alternative way of characterizing the open-loop strategy consists in noting that in correspondence of the static Cournot-Nash output we have $q'(X) = 0$. The line along which $q'(X) \rightarrow \pm\infty$ is the *non-invertibility* locus.

The lower envelope of loci $\{q^{LF}(X), \dot{X} = 0, q'(X) = 0\}$ identifies the continuum of admissible feedback solutions compatible with the non-invertibility condition. The first and last of such solutions are the linear feedback and the open-loop one. The remaining ones are the infinitely many non-linear solutions generated by the intersections between isoclines and the locus $\dot{X} = 0$, including the degenerate one at the tangency point. The arrows appearing along the isoclines (including the one tangent to $\dot{X} = 0$) reveal that all solutions along the

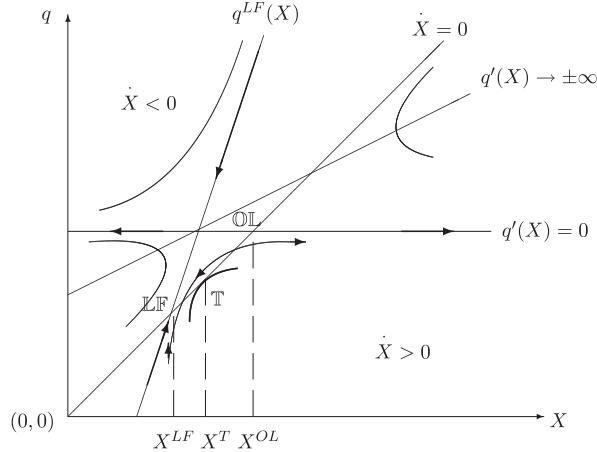


Figure 13.4 Linear and non-linear feedback solutions in the (X, q) space

segment TOL are unstable (including the extremes, points T and OL) while all those along the segment LFT (with the exception of T) are stable.

A traditional view about feedback vs open-loop strategies (in particular, in differential games based on the Cournot model) holds that since firms become more aware of their strategic interaction under feedback information, then a preemption effect takes place and outputs are larger than under open-loop information (see Fershtman and Kamien, 1987; Reynolds, 1987, 1991; Cellini and Lambertini, 2004, *inter alia*). This, in a model of resource exploitation, means bad news for the resource or species, because we must expect firms to become more aggressive in choosing their extraction plans.

However, observing Figure 13.4, there emerges that, if indeed the proper linear feedback strategy $q^{LF}(X)$ is the most aggressive of all, and this is self-evident from the residual steady state stock X^{LF} that it leaves in the sea or the ground, it is also true that there are infinitely many other (non-linear) feedback strategies less aggressive than the linear one, as firms move north-east along LFT towards T . These observations about the presence of a continuum of subgame or Markov perfect equilibria generate several question marks, e.g., as to how to regulate access to the common so as to (i) achieve a Pareto improvement and (ii) possibly shrink the set of equilibria to one.

Another delicate aspect of the above analysis is the assumption whereby $\eta > \rho(n^2 + 1)/2$, whose role is to ensure that the resource stock will not be exhausted when firms follow the feedback linear strategy, and therefore also any other type of stable strategy (those unstable may in fact annihilate the resource if its initial stock is low enough). Relaxing these assumptions, Lambertini and Mantovani (2014, pp. 119–21) investigate the viability of different strategies in terms of their long-run consequences on the resource. The qualitative properties of their analysis can be appreciated by examining Figure 13.5 (Lambertini and Mantovani, p. 120), where the steady state aggregate extraction rate is drawn against the number of firms operating in the industry.

Keeping in mind that in steady state the residual stock is proportional to aggregate output, a quick look at the curves suffices to reveal that $Q^{OL} > Q^T > Q^{LF}$ and therefore $X^{OL} > X^T > X^{LF}$ for all $n \geq 2$. Moreover, relaxing the assumption on the growth rate

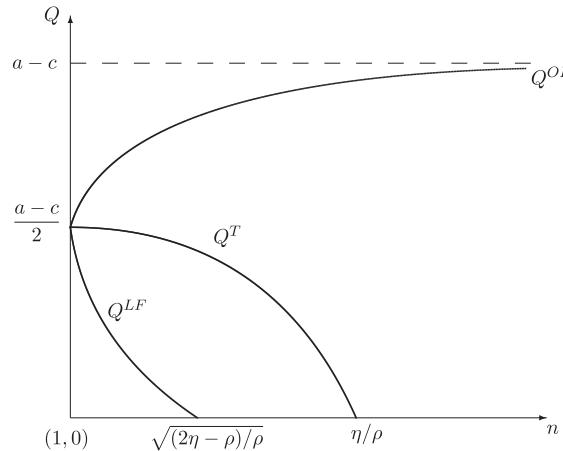


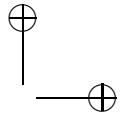
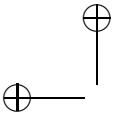
Figure 13.5 Aggregate equilibrium extraction and industry structure

entails that indeed feedback rules can compromise the preservation of the resource, if the population of firms exceeds a well-defined number. Yet, against the aforementioned intuition about the endogenous link between feedback information and output expansion, it seems that aggregate output is higher under open-loop information than under any feedback rule, except in the trivial case of monopoly. The explanation lies in two facts. The first is that the feedback effect operates throughout the game at any t , where indeed the preemption effect is visible as long as the stock is sufficiently high. The second is that the feedback effect is accompanied by the *voracity effect* (Lane and Tornell, 1996; Tornell and Lane, 1999; and Benchekroun, 2008), which can be spelled out as follows: one should not take for granted the *a priori* intuition that the higher the resource growth rate, the higher the steady state volume of that resource should be, as, in general, it may not be correct. The explanation lies in the fact that a higher level of η induces firms to hasten extraction, with the consequence that the steady state stock will be decreasing in the level of the natural growth rate of the resource. Lambertini and Mantovani (2014, pp. 121–2) illustrate the emergence of a voracity effect, and show that increasing the number of firms may make its appearance less likely under feedback rules.¹¹

A relevant extension of this approach, allowing for the analysis of Bertrand competition, is in Colombo and Labrecciosa (2015) where product differentiation is introduced à la Singh and Vives (1984).¹² This permits us to carry out a comparative evaluation of the performance of the industry in terms of profits, consumer surplus and welfare, both during the game and at the steady state, under linear and non-linear feedback strategies. The bottom line of Colombo and Labrecciosa (2015) is that, contrary to what we are accustomed to think on the basis of

¹¹ Evaluating output levels at the feedback equilibria of this game, Fujiwara (2008) concludes that oligopoly is less competitive than monopoly (see also Fujiwara, 2011). This, however, does not apply along the equilibrium path, with extraction rates being intensified by the feedback effect and the voracity effect at the same time.

¹² The same authors have investigated the linear feedback solutions of a similar game in which firms sell a homogeneous good whose production is based on the exploitation of several resource pools, each allocated to a single firm (Colombo and Labrecciosa, 2013a, 2013b). For a similar but approach to the exploitation of privately owned pools where open-loop rules are subgame perfect, see also Eswaran and Lewis (1985).



the traditional view about price vs quantity competition, here welfare is higher under quantity-setting behaviour, and under some acceptable conditions Cournot Pareto-dominates Bertrand, as both the discounted flow of profits and the discounted flow of consumer surplus are higher under Cournot behaviour than under Bertrand behaviour. The intuition for this result lies in the fact that quantity (price) strategies are increasing (decreasing) in the resource stock, and any price cuts practised by a firm today necessarily shrink the residual stock available for future harvesting by all firms alike, and the penalization for price cuts is more severe in Bertrand, due to its very nature. Consequently, price reductions (achieved by manoeuvring outputs) are more appealing to firms in the Cournot setting, which therefore tends to become more competitive than the Bertrand one.

4.1.3 The optimal number of firms in the commons

Free access to the commons is the driver of the original formulation of the tragedy in Gordon (1954) and Hardin (1968). This, in terms of a differential oligopoly games, translates into the question of whether there exists an optimal industry structure, or, an optimal number of firms in the commons. The analysis of this problem can be traced back to Cornes, Mason and Sandler (1986), Mason, Sandler and Cornes (1988) and Mason and Polasky (1997, 2002). The essence of their approach can be grasped at a glance through a summary of the model investigated in Mason and Polasky (1997).

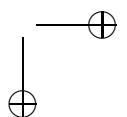
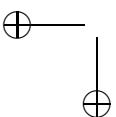
The time horizon is infinite; the natural resource follows the state equation (13.4). The market demand function $p(Q)$ is at least quasi-concave in industry output. It is exploited by n symmetric firms whose technology is summarized by the cost function $C_i(q_i, X) = [c(Q) + d(X)]q_i$, with $c'(Q) > 0$, $c''(Q) \geq 0$, $d'(X) < 0$, $d''(X) \geq 0$. The component $d(X)$ is the cost associated with resource depletion. The game is non-cooperative with each firm choosing the harvest path to maximize the discounted profit flow under the constraint given by the differential equation (13.4).

Solving for the subgame perfect equilibrium strategies of the non-cooperative game and then for the socially optimal extraction path, Mason and Polasky (1997, pp. 1153–8) characterize the modified golden rule, striking the balance between the natural resource's growth rate and time discounting and find out that the socially optimal number of firms at the steady state exceeds one for any growth rate of the resource and any level of time discounting.¹³ However, they also demonstrate that such a number will be suboptimal at least over a portion of the time horizon, or, equivalently, for some stock levels, which amounts to saying that the socially optimal number of firms in a differential game modelling the tragedy of the commons is neither time nor stock invariant and should change as time goes by and the resource depletes.

4.2 Non-renewable Resources and Cartel Behaviour

If the resource is non-renewable and $\eta = 0$, its dynamics becomes $\dot{X} = -Q$, and the problem faced by one or more firms exploiting this resource is to define an extraction plan and therefore also a pricing rule allowing for the full rent to be – almost literally – extracted before it is too late, i.e., before the technology relying on the resource in use is replaced by an alternative

¹³ This openly contradicts Hotelling's (1931) claim, according to which monopoly is the best market regime for resource preservation. See also Clemhout and Wan (1985b).



one. If the resource at stake is oil or coke, then the rent generated by these fossil fuels must be appropriated before they are replaced by some cocktail of green renewables. That is, the source of the theory of differential games involving non-renewables is in the pathbreaking paper by Hotelling (1931), whose derivation of the so-called Hotelling rule describing the intertemporal evolution of the optimal price of the resource is the outcome of an optimal control problem treated under the assumption of perfect competition. In Hotelling's own words, which are worth citing at length:

Since it is a matter of indifference to the owner whether he receives for a unit of his product [the non-renewable asset] a price p_0 now or a price $p_0 e^{\rho t}$ after time t , it is not unreasonable to expect that the price p will be a function of the time of the form $p = p_0 e^{\rho t}$. This will not apply to monopoly, where the form of the demand function is bound to affect the rate of production, but is characteristic of completely free competition. (Hotelling, 1931, p. 140)

The Hotelling model of non-renewable resource extraction is illustrated in detail elsewhere (see, e.g., Dasgupta and Heal, 1979; and Pearce and Turner, 1989), so here I will confine myself to the familiar graphical illustration of the resulting Hotelling price rule. Assuming a downward-sloping demand function $Q = a - p$ and a given stock X whose extraction is economically feasible, this reasoning generates the well-known graph reported in Figure 13.6, where p_r is the price associated with a replacement technology expected to become available at time T .¹⁴ At time T , $p_0 e^{\rho T} = p_r$ must hold and the rent generated by the asset must have been fully exploited. Hence, given a generic pair (T, p_r) , the matter reduces to an optimal control problem that must be solved to find the correct value of the initial price, p_0 .

Any perturbations in the extraction technology, the demand level a , the expected replacement date T , the price of the new technology p_r , the stock X and the discount rate ρ modify the pricing curve, possibly producing kinks in it corresponding to the specific shock taking place (see Pearce and Turner, 1989, ch. 18).¹⁵

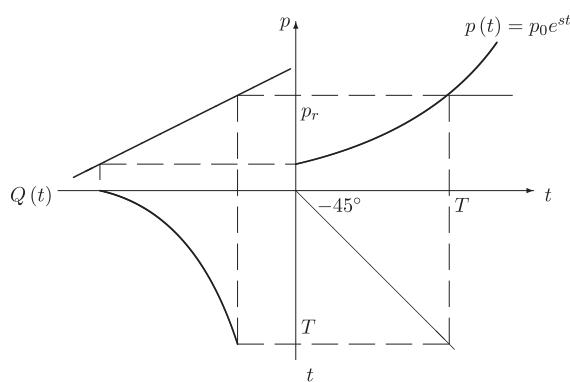


Figure 13.6 The Hotelling rule

¹⁴ For the sake of simplicity, neither the extraction cost nor the resulting royalty appear in the figure.

¹⁵ If the resource is a fossil fuel, this may also happen because of regulation or international agreements to mitigate global warming. Perturbing the pricing rule means modifying the intertemporal extraction rate. If the latter increases, the *green paradox* (Sinn, 2012) may take place. For the flourishing discussion on this issue, see Grafton, Kompas and Long (2012), Smulders, Tsur and Zemel (2012), Van der Ploeg and Withagen (2012) and Pittel, Van der Ploeg and Withagen (2014), *inter alia*.

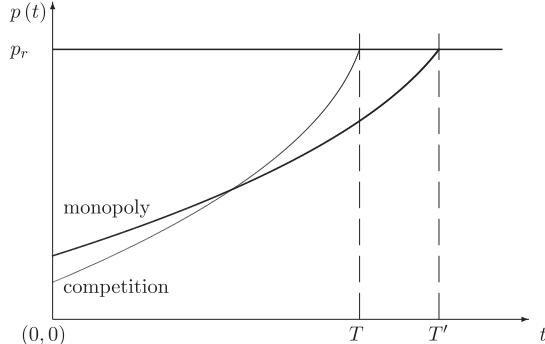


Figure 13.7 Monopoly vs perfect competition

For the purpose of the present survey, it is relevant to illustrate what happens if the market switches from perfect competition to pure monopoly, with the single firm being endowed with full control on price. A monopolistic firm raises price and shrinks production (i.e., extraction) from the initial instant. All else equal (in particular, for any given asset stock), a higher value of p_0 generates a flatter price path and consequently expands the time horizon T as compared to perfect competition. This fundamental difference between the two opposite market regimes is illustrated by the two curves in Figure 13.7. Any curve showing an intermediate pattern between them may represent the behaviour of a more or less concentrated (oligopolistic) industry in which firms do have some degree of market power and therefore exert some control on the price of the resource.

The bottom line of the analysis of the dynamic exploitation of an exhaustible resource by an industry where a cartel (at least a partial one) operates is that cartelization softens the problem of resource depletion as compared to any more competitive market regime. We know this from Salant (1976), where a dominant firm (or a cartel involving some of the firms in the industry) coexists with a competitive fringe consisting of price-taking agents. Salant (1976, p. 1085) is the first to show the presence of a two-phase extraction path in such a game: in the first phase, the price increases monotonically in the discount rate and both the cartel and the fringe extract and sell positive amounts; in the second, only the cartel supplies a positive quantity. The duration of the second phase is determined by the level of the choke price at which demand becomes nil. While Salant uses the simplifying assumption that extraction takes place at zero costs, the presence of extraction costs, being possibly asymmetric between the cartel and the fringe is accounted for in Ulph and Folie (1980). These authors and others (e.g., Gilbert, 1978; Pindyck, 1978; Lewis and Schmalensee, 1980; Newbery, 1981; Loury, 1986; Gaudet and Long, 1994) systematically adopt open-loop information.

As is well known in differential game theory (see, e.g., Dockner et al., 2000, ch. 5) the Stackelberg setup with sequential play at every instant poses a serious problem when it comes to attaining strong time consistency (i.e., subgame perfection), though obviously being in line with causal observation. This point has been tackled in the cartel vs competitive fringe model by Groot, Withagen and De Zeeuw (2003), which also maintains several essential features of the earlier open-loop literature from Salant (1976) onwards.

Consider an instantaneous demand function defined in a linear way,

$$p(t) = a - q_C(t) - Q_F(t) \quad (13.27)$$

in which subscripts \mathbb{C} and \mathbb{F} refer to the cartel and the fringe, respectively. The cartel can be thought of as a single entity (perhaps with multiple extraction plants) acting as a dominant firm; the fringe is made up of n firms, whereby $Q_{\mathbb{F}}(t) = \sum_{i=1}^n q_{\mathbb{F}i}(t)$. Extraction technologies are asymmetric: $c_{\mathbb{C}}$ and $c_{\mathbb{F}}$ identify the marginal costs associated with the cartel and each fringe firm's technologies, with $c_{\mathbb{C}}, c_{\mathbb{F}} \in [0, a]$. Initial setup costs are fixed, and therefore can be disregarded. It is assumed that access rights are well defined: the cartel has exclusive access to $X_{\mathbb{C}}(t)$ while each of the fringe members may exploit a specific pool $X_{\mathbb{F}i}(t)$. Hence, the game features $n - 1$ states and as many controls. The dynamic equations of the states are

$$\dot{X}_{\mathbb{C}} = -q_{\mathbb{C}}(t); \dot{X}_{\mathbb{F}i} = -q_{\mathbb{F}i}(t), \quad (13.28)$$

and the set of Bellman equations writes as follows:

$$\frac{\partial V_{\mathbb{C}}(\cdot)}{\partial t} + \max_{q_{\mathbb{C}}} \left[\pi_{\mathbb{C}}(\mathbf{Q}) e^{-\rho t} - \frac{\partial V_{\mathbb{C}}(\cdot)}{\partial X_{\mathbb{C}}} q_{\mathbb{C}} - \sum_{i=1}^n \frac{\partial V_{\mathbb{C}}(\cdot)}{\partial X_{\mathbb{F}i}} q_{\mathbb{F}i} \right] = 0; \quad (13.29)$$

$$\frac{\partial V_{\mathbb{F}i}(\cdot)}{\partial t} + \max_{q_{\mathbb{F}i}} \left[\pi_{\mathbb{F}i}(\mathbf{Q}) e^{-\rho t} - \frac{\partial V_{\mathbb{F}i}(\cdot)}{\partial X_{\mathbb{C}}} q_{\mathbb{C}} - \sum_{i=1}^n \frac{\partial V_{\mathbb{F}i}(\cdot)}{\partial X_{\mathbb{F}i}} q_{\mathbb{F}i} \right] = 0. \quad (13.30)$$

Solving the game à la Stackelberg, Groot et al. (2003) obtain a *modified Hotelling rule* showing that the fringe fills a demand gap (if there exists any) left by the cartel: whenever the optimal collective output of the competitive fringe is positive, the fringe itself behaves as a single price-taking agent, along the price path $p_{\mathbb{F}} = c_{\mathbb{F}} + e^{\rho t} \epsilon$, in which $\epsilon > 0$ is a parameter of the generic fringe firm's value function.

Then, Groot et al. (2003) show that the cartel's feedback strategy may give rise to four alternative regimes. The first is pure monopoly, with the fringe remaining inoperative. The second is the opposite, with perfect competition as the fringe drives the cartel out of the market. The remaining two regimes are those in which (i) the cartel uses limit pricing to exclude the fringe, or (ii) both the cartel and the fringe are active with positive market shares at equilibrium. The specific conditions on extraction costs generating each equilibrium are identified in Groot et al. (2003, pp. 292–4), with the last one emerging when costs are symmetric. The most important implication of this scenario is the resulting two-phase price path appearing in Figure 13.8, a feature that already emerged under open-loop information in Salant (1976), although with different characteristics.

The two convex curves describe the intertemporal price patterns associated with the fringe and the cartel, respectively. The horizontal line sets the choke price a . Intuitively, at any time t , the market price is identified along the lower envelope of $p_{\mathbb{C}}$ and $p_{\mathbb{F}}$. Hence, the cartel and the fringe indeed coexist with positive market shares over $t \in [0, t_1]$. Instead, the market is monopolized by the cartel for all $t \in [t_1, t_2]$. At any $t > t_2$, since $p_{\mathbb{C}}$ is larger than the choke price, the resource does not sell any more. As intuition suggests, the longer the time interval $[t_1, t_2]$ during which the cartel holds monopoly power, the lower the exploitation rate exerted by the whole industry.

Important extensions are in Benchekroun, Gaudet and Long (2006), where it is shown that the perspective of a switch to oligopolistic interaction in the future hastens extraction

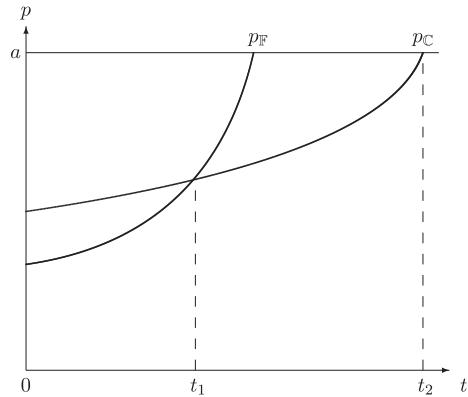


Figure 13.8 Price patterns and regime switch in the cartel vs fringe game

by the members of a temporary cartel; and in Benchekroun, Halsema and Withagen (2009, 2010), illustrating a non-cooperative game between two groups of firms differing in terms of both reserves and technological efficiency levels. Interestingly, the equilibrium of this game resembles that of the cartel vs fringe model in the limit, when the number of firms endowed with the less efficient technology becomes infinitely large.

The feedback solution of extraction games involving an exhaustible resource is rare. Early investigations in this direction can be found in Eswaran and Lewis (1985) and Reinganum and Stokey (1986), and subsequently in Salo and Tahvonen (2001). In the first of these papers, the authors compare the cases of (i) isoelastic demand function and zero extraction costs and (ii) linear demand and quadratic extraction costs, to single out the conditions whereby feedback and open-loop equilibria coincide. Reinganum and Stokey (1985) use an isoelastic demand accompanied by zero extraction costs to investigate the effect of the length of the extraction period on the pattern of the extraction rate over time. A richer setup is used in Salo and Tahvonen (2001), where each firm exploits a pool of its own, whose size is given and known, and the extraction cost is an increasing function of the depth reached by the firm itself. The most important implication of this model is that market shares tend to become symmetric as the market price increases, independently of the initial features of individual pools.

4.3 Costly Exploration and Replacement Technologies

The differential games discussed above take the initial stock of the resource as given and disregard the realistic possibility of firms drilling to expand the dimension of the pool or introduce alternative (replacement) technologies. The first perspective is accounted for in a stream of literature stemming from Peterson (1978) and including Arrow and Chang (1982), Mohr (1988) and Quyen (1988, 1991), among others. In these contributions, the drivers of exploration are the uncertainty about the size and quality of the resource deposits, the abatement of extraction costs through the expansion of the available pools or the possibility of preempting rivals by acquiring exclusive property rights on deposits still unexploited. A strategic motive of a similar nature is also formalized in Boyce and Vojtassak (2008) using a framework that traces back to Salant (1976) and Loury (1986): here, overinvestment in

exploration is practised by those firms whose certain reserves are exhausted before the same firms can certify new ones.

A direct connection with a crucial feature of the pristine Hotelling (1931) model, i.e., the prospected adoption of a new technology, appears in another flow of contributions concerning efforts to attain the so-called *backstop technologies* based on renewable resources (see Hoel, 1978; Dasgupta, Gilbert and Stiglitz, 1983; Gallini, Lewis and Ware, 1983; and Olsen, 1988). In most of these papers, the innovation date is assumed to be deterministic. Instead, Harris and Vickers (1995) consider a game where the expected date of introduction of the new technology is uncertain and depends on the intensity of the R&D effort. The search for a backstop technology relying on a renewable resource is carried out by a country importing a non-renewable one supplied by a foreign cartel. As a result, the extraction path of the resource in use can be non-monotone in its residual stock, as the cartel tries to delay the adoption of the backstop technology by the importing country, thereby modifying the price path as well, in comparison with the original Hotelling rule.

5 TWO EGGS IN ONE BASKET

Polluting emissions and natural resource exploitation should be jointly considered in a full-fledged model for at least two obvious reasons. The first is that the use of fossil fuels is a primary source of carbon emissions and global warming, i.e., exhaustible resources are typically brown. The second is that polluting emissions affect the growth rate of renewable resources and natural species, both directly and indirectly via the induced changes in the planet's climate. Alas, the available literature treating pollution and natural resources in a unique theoretical setup is still quite small.

The earliest example of this kind is in Wirl (1994, 1995), where flow and stock externalities (acid rain and global warming) are generated by cartelized fossil fuels and a tax is levied on emissions.¹⁶ Comparing linear and non-linear feedback strategies, Wirl finds that the latter are Pareto-inferior as compared to the former, so that firms, consumers and the policy maker would be happy to confine the solution of the game to linear feedback rules. Wirl and Dockner (1995) modify the model to allow for the presence of a Leviathan government attaching a value to tax revenue per se, in addition to the traditional notion of welfare.

A further extension of the same model (Tahvonen, 1996) examines Stackelberg play, with the cartel taking the lead. It turns out that the efficient tax is higher than that associated with Stackelberg play. This implies that the cartel has the ability of shaping the price pattern in order to neutralize or at least soften pressure exerted by the tax. Rubio and Escriche (2001) highlight the possibility of using the tax as a dual instrument, so as to diminish the externality (its declared aim) and extract a fraction of the cartel's rent.¹⁷

¹⁶ Most of this literature assumes cartel behaviour by the extracting industry. For an analysis of non-cooperative Cournot behaviour in models where open-loop strategies are subgame perfect, see Dragone, Lambertini, Palestini and Tampieri (2013) and Lambertini and Leitmann (2013).

¹⁷ In this respect, it is worth mentioning that several studies stemming from Ulph and Ulph (1994) discuss the relationship between the optimal carbon tax and the residual stock of the exhaustible (fossil) resource being exploited. Ulph and Ulph (1994) show that the optimal tax should decrease over time as the residual resource stock decreases. Yet, if the extraction rate is not an appropriate measure of emissions, the optimal tax pattern is concave over time and therefore non-monotone in the residual stock. For more, see Amundsen and Schöb (1999); Liski and Tahvonen (2004); Wirl (2007); Wei et al. (2012); and Prieur, Tidball and Withagen (2013).

6 CSR AND THE PORTER HYPOTHESIS

Here I will review a few contributions incorporating two elements that are currently attracting a large amount of attention: *corporate social responsibility* (CSR) and the *Porter hypothesis*. Both are being taken attentively into consideration as two distinctive features of a sustainable firm behaviour.

CSR is a wide-ranging label encompassing a number of good practices whereby a firm could considerably diminish its environmental impact. Indeed, it can be viewed as a form of corporate self-regulation, and therefore the possibility of its spontaneous adoption by firms has been animatedly questioned (see Baron, 2001, 2007; and Benabou and Tirole, 2010). Nonetheless, recent research in this direction points out the possibility for CSR to enhance a firm's profitability, at least in mixed markets where competitors stick to a pure profit-seeking behaviour (see, e.g., Lambertini and Tampieri, 2015).¹⁸

This result has been reproduced in a differential game setting à la Ramsey (1928) by Lambertini, Palestini and Tampieri (2016) considering a Cournot duopoly in which firms supply a homogeneous good whose demand function is $p(t) = a - q_1(t) - q_2(t)$, sharing the same technology characterized by a constant marginal production cost $c \in (0, a)$. Each firm uses a linear production function $y_i(t) = Ak_i(t)$, and accumulates productive capacity $k_i(t)$ according to the following state equation:

$$\dot{k}_i = Ak_i(t) - q_i(t) - \mu k_i(t), \quad (13.31)$$

whereby the individual capital stock accumulates via unsold output and $\mu > 0$ is its constant decay rate. Production is brown, and the authors stipulate that firm i is a CSR unit, while firm 2 is a pure profit-seeking one. Obeying its CSR nature, firm i also carries out a green R&D effort $x_i(t)$ to reduce polluting emissions $S(t)$, which evolve according to the state dynamics

$$\dot{S} = q_1(t) + q_2(t) - \delta S(t) - \nu x_1(t), \quad (13.32)$$

where parameter ν measures the instantaneous effectiveness of R&D. At any t , the R&D entails a quadratic cost so that the CSR firm's profit function is $\pi_i(t) = [p(t) - c]q_i(t) - zx_i^2(t)$, with $z > 0$. The CSR firm's objective function is

$$\Omega_1(t) = \pi_1(t) - bS(t) + \gamma CS(t), \quad (13.33)$$

with $b, \gamma > 0$ representing the weights attached to the volume of emissions and consumer surplus. Solving the resulting linear state game under open-loop rules, Lambertini et al. (2016) find out plausible conditions under which the CSR firm sells more, accumulates a larger capacity and gets higher profits than its profit-seeking rival, and contributes more than its profit-seeking rival, and may deliver a welfare increase.

¹⁸ Investigating a differential oligopoly game where all firms undertake costly investments to expand their CSR stances, Wirl, Feichtinger and Kort (2013) prove the existence of multiple equilibria, whose number may either increase or decrease via the interaction among firms. The role of interaction can be as relevant as to create a situation where history is insufficient to determine the selection of one out of the many equally plausible steady states, so that coordination is indeed beneficial.

The Porter hypothesis (Porter, 1991; Porter and Van der Linde, 1995), an informal claim according to which environmental regulation may indeed produce a win-win solution whereby firms attain higher profits by going green and welfare increases as a result of firms' investment, openly challenges the acquired view that holds that any type of regulation should hinder firms' profit performance, and has duly attracted a large amount of attention by researchers in the field of environmental economics,¹⁹ including a valuable formulation via a differential game approach in which the win-win solution emerges as a consequence of firms' (dis)investment in capital inputs of different vintages and productive efficiency, proposed by Xepapadeas and De Zeeuw (1999). In a nutshell, their formalization of the Porter hypothesis relies on demonstrating that updating (and possibly shrinking) the capital endowment installed in any given firm is conducive to a reduction in polluting emissions and therefore also in the tax pressure the same firm has to bear.

For the sake of simplicity, Xepapadeas and De Zeeuw (1999) assume market price to be exogenously fixed at $p > 0$ throughout the infinite time horizon of the game. As in Feichtinger et al. (2016), this could be due to either a price cap (possibly unrelated to the environmental impact of the industry at stake) or the presence of a regulated access to emission quotas.

At the cost of eliminating a conspicuous amount of strategic interaction, this simplifying assumption allows one to take the standpoint of a single firm and treat the model as a single-agent optimization problem. The generic firm invests in capital inputs of vintage $\omega \in [0, \bar{\omega}]$. Let $q(\omega)$ be the output level produced by the technology using an input of vintage ω , with $\partial q(\omega) / \partial \omega \leq 0$. The operative cost of using this input is $c(\omega)$, with $\partial c(y) / \partial y \geq 0$; while the environmental impact of the same vintage is $S(\omega)$, with $\partial s(\omega) / \partial \omega \leq 0$. The picture implied by these features is one in which old capital inputs produce less and pollute more than new ones, at higher marginal production costs. The instantaneous output produced by the firm is a function of the spectrum of vintages in use within the same firm:

$$Q(t) = \int_0^{\bar{\omega}} q(\omega) n(\omega, t) d\omega, \quad (13.34)$$

where $n(\omega, t) \geq 0$ is the number of inputs of the same vintage ω at any t .

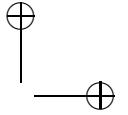
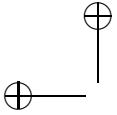
Capital inputs are tradeable. Trading $\tilde{n}(\omega, t)$ generates revenues (for the seller; costs, for the buyer) $R(\tilde{n}) = b(\omega)\tilde{n}(\cdot) + \tilde{n}(\cdot)/2$, with $b(\omega) > 0$ for all $\omega \in [0, \bar{\omega}]$, $\partial b(\omega) / \partial \omega \leq 0$ and $b(\bar{\omega}) = 0$. This amounts to saying that an input becomes cheaper with age, and the oldest vintage is indeed worthless. In presence of a tax τ levied on emissions, and the firm sells $\tilde{n}(\omega, t)$, its instantaneous profit is

$$\pi(\omega, t) = \int_0^{\bar{y}} \{[pq(t) - c(\omega) - \tau s(\omega)]n(\omega, t) - R(\tilde{n})\} dy \quad (13.35)$$

to be maximized under the state equation

$$\dot{n} = \tilde{n}(\omega, t) - \frac{\partial n(\omega, t)}{\partial \omega} \quad (13.36)$$

¹⁹ The resulting literature is already too large to be exhaustively listed here. For a comprehensive survey, see Ambec et al. (2013).



and the initial condition $n(0,0) = 0$. The differential equation (13.36) describes the evolution of the vintage composition of capital installed in the firm in terms of trade and breakdowns due to ageing. The firm has to maximize Hamiltonian function

$$\mathcal{H} = \pi(\omega, t) + \lambda(t) \left[\tilde{n}(\omega, t) - \frac{\partial n(\omega, t)}{\partial \omega} \right] \quad (13.37)$$

using the traded amount $\tilde{n}(\omega, t)$ as its control variable. Manipulating the resulting set of necessary conditions, Xepapadeas and De Zeeuw (1999) identify the optimal trade volume $\tilde{n}^* = \lambda - b$ and the optimal endowment of inputs of vintage ω :

$$n^*(\omega) = \int_0^\omega \left[\int_v^{\bar{\omega}} (pq(t) - c(t) - \tau s(t)) dt - b(v) \right] dv \quad (13.38)$$

with $\partial n^*(\omega) / \partial \tau < 0$. Looking at total emissions

$$S(\omega) = \int_v^{\bar{\omega}} s(t) n^*(\omega) dt, \quad (13.39)$$

they also find that $\partial S(\omega) / \partial \tau < 0$. That is, taxing emissions induces the firm to downsize its capital stock and consequently implies a reduction in the environmental damage generated by the technology in use. The emergence of the win-win solution confirming the Porter hypothesis is verified if the ratio measuring the average vintage of the equilibrium capital stock, i.e.,

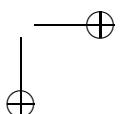
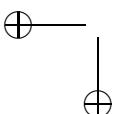
$$\text{avg}_\omega = \frac{\int_v^{\bar{\omega}} \omega n^*(\omega) d\omega}{\int_v^{\bar{\omega}} n^*(\omega) d\omega} \quad (13.40)$$

is decreasing in the emission tax, i.e., if $\partial \text{avg}_\omega(\tau) / \partial \tau < 0$: as the environmental policy becomes tighter, the average age of the spectrum of inputs used by the firm decreases. The equilibrium profits increase because shrinking the volume of emissions by refreshing the capital stock diminishes the tax bill although the latter is implemented by a higher tax rate applied to every unit of pollutants.

7 INTERNATIONAL TRADE

A comparatively smaller debate also exists about the interplay between international trade in oligopolistic industries and environmental issues, the relevance of this matter in a globalized economic system notwithstanding.²⁰ Here, I will briefly summarize the contributions of Feenstra et al. (1996), Feenstra, Kort and De Zeeuw (2001), Fujiwara (2009), Yanase (2007, 2012) and Jinji (2013).

²⁰ The opposite is true in the economics of natural resources and the environment, where this matter is being energetically debated. See, for an exhaustive account, Copeland and Taylor (2003, 2004, 2009).



Feenstra et al. (1996) evaluate the effectiveness of emission taxations vs environmental standards under intraindustry trade in a game where firms adopt open-loop strategies, to confirm a result generated in (static) multistage games, according to which standards outperform taxes (see, e.g., Ulph, 1992). Since open-loop strategies call for commitment devices while feedback strategies – being state-dependent – are more flexible, this conclusion is not necessarily reliable. To prove this, Feenstra et al. (2001) lay out a duopoly game with firms using capital and a polluting input under intraindustry trade and then perform the same comparison between the two policies under feedback rules. Doing so, they obtain a richer picture in which one can identify the parametric conditions concerning the discount factor and the marginal productivity of capital under which the emission tax outperforms the standard, or the opposite.

Fujiwara (2009) adapts the Cournot model of Benckroun and Long (1998) to model the environmental impact of trade liberalization in a two-country model with transboundary pollution and a single firm based in each country. Open-loop as well as linear and non-linear feedback strategies are characterized. Production emits pollutants involving a quadratic instantaneous environmental damage. The free trade equilibrium is compared with autarky, to find that if the efficiency of natural carbon sinks is too low, then international trade is certainly harmful because the beneficial price effect associated with the industry output expansion after the opening of trade cannot counterbalance the parallel increase in polluting emissions, which provokes a more than proportionate increase in the environmental damage.

In a similar model, Yanase (2007) compares the effects of emission taxes and emission quotas, finding out that under feedback information the emission tax regime causes a larger strategic distortion between firms (and countries) than the emission quota regime. Instead, the performance of the two regimes coincide if governments coordinate their respective policies.

Yanase (2012) and Jinji (2013) insert CSR into the general picture, in two different ways. In Yanase (2012), intraindustry trade takes place among a large number of countries in which firms have adopted corporate environmentalism. From the analysis carried out under linear feedback strategies, it emerges that in the short run (i) free trade unambiguously improves any country's welfare in the fully symmetric case, while the outcome depends on parameters if countries are asymmetric. The long-run effect is ambiguous in both cases. Jinji (2013) allows for the presence of both emission taxes and export subsidies, showing that the adoption of a CSR stance can indeed bring about a reduction in domestic welfare if pollution is transboundary. If export subsidies are ruled out, then a welfare decrease may obtain even when pollution is country-specific.

REFERENCES

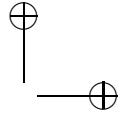
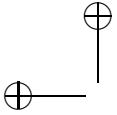
- Aghion, P., N. Bloom and R. Blundell et al. (2005), "Competition and Innovation: An Inverted-U Relationship", *Quarterly Journal of Economics*, **120**, 701–28.
 Ambec, S., M.A. Cohen, S. Elgie and P. Lanoie (2013), "The Porter Hypothesis at 20: Can Environmental Regulation Enhance Innovation and Competitiveness?", *Review of Environmental Economics and Policy*, **7**, 2–22.
 Amundsen, E.S. and R. Schöb (1999), "Environmental Taxes on Exhaustible Resources", *European Journal of Political Economy*, **15**, 311–29.
 Arrow, K.J. (1962), "Economic Welfare and the Allocation of Resources for Invention", in R. Nelson (ed.), *The Rate and Direction of Industrial Activity*, Princeton, NJ: Princeton University Press.
 Arrow, K.J. and S. Chang (1982), "Optimal Pricing, Use, and Exploration of Uncertain Resource Stocks", *Journal of Environmental Economics and Management*, **9**, 1–10.

- Baron, D. (2001), "Private Politics, Corporate Social Responsibility, and Integrated Strategy", *Journal of Economics and Management Strategy*, **10**, 7–45.
- Baron, D. (2007), "Corporate Social Responsibility and Social Entrepreneurship", *Journal of Economics and Management Strategy*, **16**, 683–717.
- Benabou, R. and J. Tirole (2010), "Individual and Corporate Social Responsibility", *Economica*, **77**, 1–19.
- Bencheikroun, H. (2003), "Unilateral Production Restrictions in a Dynamic Duopoly", *Journal of Economic Theory*, **111**, 214–39.
- Bencheikroun, H. (2008), "Comparative Dynamics in a Productive Asset Oligopoly", *Journal of Economic Theory*, **138**, 237–61.
- Bencheikroun, H. and A.R. Chaudhury (2011), "Environmental Policy and Stable Collusion: The Case of a Dynamic Polluting Oligopoly", *Journal of Economic Dynamics and Control*, **35**, 479–90.
- Bencheikroun, H. and N.V. Long (1998), "Efficiency Inducing Taxation for Polluting Oligopolists", *Journal of Public Economics*, **70**, 325–42.
- Bencheikroun, H. and N.V. Long (2002a), "On the Multiplicity of Efficiency-inducing Tax Rules", *Economics Letters*, **76**, 331–6.
- Bencheikroun, H. and N.V. Long (2002b), "Transboundary Fishery: A Differential Game Model", *Economica*, **69**, 207–21.
- Bencheikroun, H., G. Gaudet and N.V. Long (2006), "Temporary Natural Resource Cartels", *Journal of Environmental Economics and Management*, **52**, 663–74.
- Bencheikroun, H., A. Hulsema and C. Withagen (2009), "On Non-renewable Resource Oligopolies: The Asymmetric Case", *Journal of Economic Dynamics and Control*, **33**, 1867–79.
- Bencheikroun, H., A. Hulsema and C. Withagen (2010), "When Additional Resource Stocks Reduce Welfare", *Journal of Environmental Economics and Management*, **59**, 109–14.
- Benhabib, J. and R. Radner (1992), "The Joint Exploitation of Productive Asset: A Game-theoretic Approach", *Economic Theory*, **2**, 155–90.
- Bergstrom, T., J. Cross and R. Porter (1981), "Efficiency-inducing Taxation for a Monopolistically Supplied Depletable Resource", *Journal of Public Economics*, **15**, 23–32.
- Boyce, J. and L. Vojtassak (2008), "An 'Oil'igopoly Theory of Exploration", *Resource and Energy Economics*, **30**, 428–54.
- Brander, J. and S.M. Taylor (1997), "International Trade and Open Access Renewable Resources: The Small Open Economy Case", *Canadian Journal of Economics*, **30**, 526–52.
- Cave, J. (1987), "Long-term Competition in a Dynamic Game: The Cold Fish War", *RAND Journal of Economics*, **18**, 596–610.
- Cellini, R. and L. Lambertini (2004), "Dynamic Oligopoly with Sticky Prices: Closed-loop, Feedback and Open-loop Solutions", *Journal of Dynamical and Control Systems*, **10**, 303–14.
- Cellini, R., L. Lambertini and G. Leitmann (2005), "Degenerate Feedback and Time Consistency in Differential Games", in E.P. Hofer and E. Reithmeier (eds), *Modeling and Control of Autonomous Decision Support Based Systems. Proceedings of the 13th International Workshop on Dynamics and Control*, Aachen: Shaker Verlag.
- Chiarella, C., M. Kemp, N.V. Long and K. Okuguchi (1984), "On the Economics of International Fisheries", *International Economic Review*, **25**, 85–92.
- Chichilnisky, G. (1994), "North–South Trade and the Global Environment", *American Economic Review*, **84**, 851–74.
- Clark, C.W. (1973), "Profit Maximization and the Extinction of Animal Species", *Journal of Political Economy*, **81**, 950–60.
- Clark, C.W. (1990). *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*, New York: Wiley.
- Clark, C.W. and G.R. Munro (1975), "The Economics of Fishing and Modern Capital Theory", *Journal of Environmental Economics and Management*, **2**, 92–106.
- Clemhout, S. and H. Wan, Jr. (1985a), "Dynamic Common Property Resources and Environmental Problems", *Journal of Optimization Theory and Applications*, **46**, 471–81.
- Clemhout, S. and H.Y. Wan, Jr. (1985b), "Cartelization Conserves Endangered Species", in G. Feichtinger (ed.), *Optimal Control Theory and Economic Analysis*, Vol. 2, Amsterdam: North-Holland.
- Colombo, L. and P. Labrecciosa (2013a), "Oligopoly Exploitation of a Private Property Productive Asset", *Journal of Economic Dynamics and Control*, **37**, 838–53.
- Colombo, L. and P. Labrecciosa (2013b), "On the Convergence to the Cournot Equilibrium in a Productive Asset Oligopoly", *Journal of Mathematical Economics*, **49**, 441–5.
- Colombo, L. and P. Labrecciosa (2015), "On the Markovian Efficiency of Bertrand and Cournot Equilibria", *Journal of Economic Theory*, **155**, 332–58.
- Copeland, B.R. and M.S. Taylor (2003), *Trade and the Environment: Theory and Evidence*, Princeton, NJ: Princeton University Press.

- Copeland, B.R. and M.S. Taylor (2004), "Trade, Growth, and the Environment", *Journal of Economic Literature*, **42**, 7–71.
- Copeland, B.R. and S.M. Taylor (2009), "Trade, Tragedy, and the Commons", *American Economic Review*, **99**, 725–49.
- Cornes, R., C.F. Mason and T. Sandler (1986), "The Commons and the Optimal Number of Firms", *Quarterly Journal of Economics*, **101**, 641–6.
- d'Aspremont, C., A. Jacquemin, J.J. Gabszewicz and J.A. Weymark (1983), "On the Stability of Collusive Price Leadership", *Canadian Journal of Economics*, **16**, 17–25.
- Dasgupta, P.S. and G.M. Heal (1979), *Economic Theory and Exhaustible Resources*, Cambridge, UK: Cambridge University Press.
- Dasgupta, P., R. Gilbert and J. Stiglitz (1983), "Strategic Considerations in Invention and Innovation: The Case of Natural Resources", *Econometrica*, **51**, 1439–48.
- Dawid, H. and M. Kopel (1997), "On the Economically Optimal Exploitation of a Renewable Resource: The Case of a Convex Environment and a Convex Return Function", *Journal of Economic Theory*, **76**, 272–97.
- Dockner, E.J., G. Feichtinger and S. Jørgensen (1985), "Tractable Classes of Nonzero-sum Open-loop Nash Differential Games: Theory and Examples", *Journal of Optimization Theory and Applications*, **45**, 179–97.
- Dockner, E.J. and G. Sorger (1996), "Existence and Properties of Equilibria for a Dynamic Game on Productive Assets", *Journal of Economic Theory*, **71**, 201–27.
- Dockner, E.J., S. Jørgensen, N.V. Long, and G. Sorger (2000), *Differential Games in Economics and Management Science*, Cambridge, UK: Cambridge University Press.
- Donsimoni, M.P. (1985), "Stable Heterogeneous Cartels", *International Journal of Industrial Organization*, **3**, 451–67.
- Donsimoni, M.P., N.S. Economides and H.N. Polemarchakis (1986), "Stable Cartels", *International Economic Review*, **27**, 317–27.
- Dragone, D., L. Lambertini and A. Palestini (2010), "Dynamic Oligopoly with Capital Accumulation and Environmental Externality", in J. Crespo Cuaresma, T. Palokangas and A. Tarashev (eds), *Dynamic Systems, Economic Growth, and the Environment*, Heidelberg: Springer.
- Dragone, D., L. Lambertini and A. Palestini (2013), "The Incentive to Invest in Environmental-friendly Technologies: Dynamics Makes a Difference", in J. Crespo Cuaresma, T. Palokangas and A. Tarashev (eds), *Green Growth and Sustainable Development*, Heidelberg: Springer.
- Dragone, D., L. Lambertini, A. Palestini and A. Tampieri (2013), "On the Optimal Number of Firms in the Commons: Cournot vs Bertrand", *Mathematical Economics Letters*, **1**, 25–34.
- Dragone, D., L. Lambertini and A. Palestini (2014), "Regulating Environmental Externalities through Public Firms: A Differential Game", *Strategic Behavior and the Environment*, **4**, 15–40.
- Eswaran, M. and T. Lewis (1985), "Exhaustible Resources and Alternative Equilibrium Concepts", *Canadian Journal of Economics*, **18**, 459–73.
- Feenstra, T., P. Kort and A. de Zeeuw (2001), "Environmental Policy Instruments in an International Duopoly with Feedback Investment Strategies", *Journal of Economic Dynamics and Control*, **25**, 1665–87.
- Feenstra, T., P. Kort, P. Verheyen and A. de Zeeuw (1996), "Standards versus Taxes in a Dynamic Duopoly Model of Trade", in A. Xepapadeas (ed.), *Economic Policy for the Environment and Natural Resources*, Cheltenham, UK and Brookfield, VT, USA: Edward Elgar Publishing.
- Feichtinger, G., L. Lambertini, G. Leitmann and S. Wrzaczek (2016), "R&D for Green Technologies in a Dynamic Oligopoly: Schumpeter, Arrow and Inverted U's", *European Journal of Operational Research*, **249**, 1131–8.
- Fershtman, C. (1987), "Identification of Classes of Differential Games for Which the Open-loop is a Degenerated Feedback Nash Equilibrium", *Journal of Optimization Theory and Applications*, **55**, 217–31.
- Fershtman, C. and M. Kamien (1987), "Dynamic Duopolistic Competition with Sticky Prices", *Econometrica*, **55**, 1151–64.
- Fischer, R.D. and L. Mirman (1986), "The Compleat Fish Wars: Biological and Dynamic Interactions", *Journal of Environmental Economics and Management*, **30**, 34–42.
- Fujiwara, K. (2008), "Duopoly Can Be More Anti-competitive Than Monopoly", *Economics Letters*, **101**, 217–19.
- Fujiwara, K. (2009), "Why Environmentalists Resist Trade Liberalization", *Environmental and Resource Economics*, **44**, 71–84.
- Fujiwara, K. (2011), "Losses from Competition in a Dynamic Game Model of a Renewable Resource Oligopoly", *Resource and Energy Economics*, **33**, 1–11.
- Gallini, N., T. Lewis and R. Ware (1983), "Strategic Timing and Pricing of a Substitute in a Cartelized Resource Market", *Canadian Journal of Economics*, **16**, 429–46.
- Gaudet, G. and N.V. Long (1994), "On the Effects of the Distribution of Initial Endowments in a Non-renewable Resource Duopoly", *Journal of Economic Dynamics and Control*, **18**, 1189–98.
- Gilbert, R. (1978), "Dominant Firm Pricing Policy in a Market for an Exhaustible Resource", *Bell Journal of Economics*, **9**, 385–95.

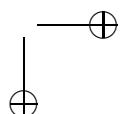
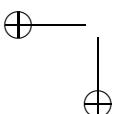
- Gordon, H.S. (1954), "The Economic Theory of a Common-property Resource: The Fishery", *Journal of Political Economy*, **62**, 124–42.
- Grafton, R.Q., T. Kompas and N.V. Long (2012), "Substitution between Biofuels and Fossil Fuels: Is there a Green Paradox?", *Journal of Environmental Economics and Management*, **64**, 328–41.
- Groot, F., C. Withagen and A. de Zeeuw (2003), "Strong Time-consistency in the Cartel-versus-Fringe Model", *Journal of Economic Dynamics and Control*, **28**, 287–306.
- Hardin, G. (1968). "The Tragedy of the Commons", *Science*, **162**, 1243–8.
- Harris, C. and J. Vickers (1995), "Innovation and Natural Resources: A Dynamic Game with Uncertainty", *RAND Journal of Economics*, **26**, 418–30.
- Hoel, M. (1978), "Resource Extraction, Substitute Production, and Monopoly", *Journal of Economic Theory*, **19**, 28–77.
- Hotelling, H. (1931), "The Economics of Exhaustible Resources", *Journal of Political Economy*, **39**, 137–75.
- Jinji, N. (2013), "Is Corporate Environmentalism Good for Domestic Welfare?", *Review of International Economics*, **21**, 901–11.
- Jørgensen, S., G. Martín-Herrán and G. Zaccour (2010), "Dynamic Games in the Economics and Management of Pollution", *Environmental Modelling and Assessment*, **15**, 433–67.
- Karp, L. and J. Livernois (1994), "Using Automatic Tax Changes to Control Pollution Emissions", *Journal of Environmental Economics and Management*, **27**, 38–48.
- Lambertini, L. (2013), *Oligopoly, the Environment and Natural Resources*, London: Routledge.
- Lambertini, L. (2016), "Managerial Delegation in a Dynamic Renewable Resource Oligopoly", in H. Dawid, K. Doerner and G. Feichtinger et al. (eds), *Dynamic Perspectives on Managerial Decision Making: Essays in Honor of Richard F. Hartl*, Heidelberg: Springer.
- Lambertini, L. and G. Leitmann (2013), "Market Power, Resource Extraction and Pollution: Some Paradoxes and a Unified View", in J. Crespo Cuaresma, T. Palokangas and A. Tarashev (eds), *Green Growth and Sustainable Development*, Heidelberg: Springer.
- Lambertini, L. and A. Mantovani (2014), "Feedback Equilibria in a Dynamic Renewable Resource Oligopoly: Pre-emption, Voracity and Exhaustion", *Journal of Economic Dynamics and Control*, **47**, 115–22.
- Lambertini, L. and A. Mantovani (2016), "On the (In)stability of Non-linear Feedback Solutions in a Dynamic Duopoly with Renewable Resource Exploitation", *Economics Letters*, **143**, 9–12.
- Lambertini, L. and A. Tampieri (2015), "Incentive, Performance and Desirability of Socially Responsible Firms in a Cournot Oligopoly", *Economic Modelling*, **50**, 40–48.
- Lambertini, L., A. Palestini and A. Tampieri (2016), "CSR in Asymmetric Duopoly with Environmental Externality", *Southern Economic Journal*, **83**, 236–52.
- Lane, P.R. and A. Tornell (1996), "Power, Growth, and the Voracity Effect", *Journal of Economic Growth*, **1**, 213–41.
- Leibenstein, H. (1966), "Allocative Efficiency versus X-efficiency", *American Economic Review*, **56**, 392–415.
- Levhari, D. and L. Mirman (1980), "The Great Fish War: An Example Using a Dynamic Cournot-Nash Solution", *Bell Journal of Economics*, **11**, 322–34.
- Lewis, T. and R. Schmalensee (1980), "Cartel and Oligopoly Pricing of Non-replenishable Natural Resources", in P. Liu (ed.), *Dynamic Optimization and Application to Economics*, New York: Plenum Press.
- Liski, M. and O. Tahvonen (2004), "Can Carbon Tax Eat OPEC's Rents?", *Journal of Environmental Economics and Management*, **47**, 1–12.
- Long, N.V. (2010), *A Survey of Dynamic Games in Economics*, Singapore: World Scientific.
- Long, N.V. (2011), "Dynamic Games in the Economics of Natural Resources: A Survey", *Dynamic Games and Applications*, **1**, 115–48.
- Lotka, A.J. (1925), *Elements of Physical Biology*, Philadelphia, PA: Williams and Wilkins.
- Loury, G. (1986), "A Theory of 'Oil'igopoly: Cournot Equilibrium in Exhaustible Resource Market with Fixed Supplies", *International Economic Review*, **27**, 285–301.
- Mason, C.F. and S. Polasky (1997), "The Optimal Number of Firms in the Commons: A Dynamic Approach", *Canadian Journal of Economics*, **30**, 1143–60.
- Mason, C. and S. Polasky (2002), "Strategic Preemption in a Common Property Resource: A Continuous Time Approach", *Environmental and Resource Economics*, **23**, 255–78.
- Mason, C., T. Sandler and R. Cornes (1988), "Expectations, the Commons, and Optimal Group Size", *Journal of Environmental Economics and Management*, **15**, 99–110.
- Mehlmann, A. (1988), *Applied Differential Games*, New York: Plenum Press.
- Mohr, E. (1988), "Appropriation of Common Access Natural Resources Through Exploration: The Relevance of the Open-loop Concept", *International Economic Review*, **29**, 307–20.
- Newbery, D. (1981), "Oil Prices, Cartels, and the Problem of Dynamic Inconsistency", *Economic Journal*, **91**, 617–46.
- Olsen, T.E. (1988), "Strategic Considerations in Invention and Innovation: The Case of Natural Resources Revisited", *Econometrica*, **56**, 841–9.

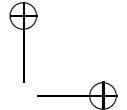
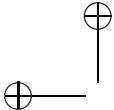
- Pearce, D.W. and R.K. Turner (1989), *Economics of Natural Resources and the Environment*, Hemel Hempstead, UK: Harvester-Wheatsheaf.
- Peterson, F.M. (1978), "A Model of Mining and Exploration for Exhaustible Resources", *Journal of Environmental Studies and Management*, **5**, 236–51.
- Pindyck, R. (1978), "Gains to Producers from the Cartelization of Exhaustible Resources", *Review of Economics and Statistics*, **60**, 238–51.
- Pittel, K. F. van der Ploeg and C. Withagen (2014), *Climate Policy and Non-renewable Resources: The Green Paradox and Beyond*, Cambridge, MA: MIT Press.
- Porter, M. (1991), "America's Green Strategy", *Scientific American*, **264**, 168.
- Porter, M. and C. van der Linde (1995), "Toward a New Conception of the Environment–competitiveness Relationship", *Journal of Economic Perspectives*, **9**, 97–118.
- Prieur, F., M. Tidball and C. Withagen (2013), "Optimal Emission-extraction Policy in a World of Scarcity and Irreversibility", *Resource and Energy Economics*, **35**, 637–58.
- Quyen, N.V. (1988), "The Optimal Depletion and Exploration of a Non-renewable Resource", *Econometrica*, **56**, 1467–71.
- Quyen, N.V. (1991), "Exhaustible Resources: A Theory of Exploration", *Review of Economic Studies*, **58**, 777–89.
- Ramsey, F.P. (1928), "A Mathematical Theory of Saving", *Economic Journal*, **38**, 543–59.
- Reinganum, J. and N. Stokey (1985), "Oligopoly Extraction of a Common Property Resource: The Importance of the Period of Commitment in Dynamic Games", *International Economic Review*, **26**, 161–73.
- Requate, T. (2005), "Dynamics Incentives by Environmental Policy Instruments – A Survey", *Ecological Economics*, **54**, 175–95.
- Reynolds, S. (1987), "Preemption and Commitment in an Infinite Horizon Model", *International Economic Review*, **28**, 69–88.
- Reynolds, S. (1991), "Dynamic Oligopoly with Capacity Adjustment Costs", *Journal of Economic Dynamics and Control*, **15**, 491–514.
- Rowat, C. (2007), "Non-linear Strategies in a Linear Quadratic Differential Game", *Journal of Economic Dynamics and Control*, **31**, 3179–202.
- Rubio, S. and L. Escriche (2001), "Strategic Pigouvian Taxation, Stock Externalities and Non-renewable Resources", *Journal of Public Economics*, **79**, 297–313.
- Salant, S.W. (1976), "Exhaustible Resources and Industrial Structure: A Nash-Cournot Approach to the World Oil Market", *Journal of Political Economy*, **84**, 1079–94.
- Salo, S. and O. Tahvonen (2001), "Oligopoly Equilibria in Non-renewable Resource Markets", *Journal of Economic Dynamics and Control*, **25**, 671–702.
- Schumpeter, J.A. (1942), *Capitalism, Socialism and Democracy*, London: Allen & Unwin.
- Singh, N. and X. Vives (1984), "Price and Quantity Competition in a Differentiated Duopoly", *RAND Journal of Economics*, **15**, 546–54.
- Sinn, H.W. (2012), *The Green Paradox*, Cambridge, MA: MIT Press.
- Smulders, S., Y. Tsur and A. Zemel (2012), "Announcing Climate Policy: Can a Green Paradox Arise without Scarcity?", *Journal of Environmental Economics and Management*, **64**, 364–76.
- Tahvonen, O. (1996), "Trade with Polluting Non-renewable Resources", *Journal of Environmental Economics and Management*, **30**, 1–17.
- Tornell, A. and P.R. Lane (1999), "The Voracity Effect", *American Economic Review*, **89**, 22–46.
- Tsutsui, S. and K. Mino (1990), "Non-linear Strategies in Dynamic Duopolistic Competition with Sticky Prices", *Journal of Economic Theory*, **52**, 136–61.
- Ulph, A. (1992), "The Choice of Environmental Policy Instruments and Strategic International Trade", in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources*, Heidelberg: Springer.
- Ulph, A. and G.M. Folie (1980), "Exhaustible Resources and Cartels: An Intertemporal Nash-Cournot Model", *Canadian Journal of Economics*, **13**, 645–58.
- Ulph, A. and D. Ulph (1994), "The Optimal Time Path of a Carbon Tax", *Oxford Economic Papers*, **46**, 857–68.
- Van der Ploeg, F. and C. Withagen (2012), "Is There Really a Green Paradox?", *Journal of Environmental Economics and Management*, **64**, 342–63.
- Verhulst, P.H. (1838), "Notice sur la loi que la population poursuit dans son accroissement", *Correspondance mathématique et physique*, **10**, 113–21.
- Volterra, V. (1931), "Variations and Fluctuations of the Number of Individuals in Animal Species Living Together", in R.N. Chapman (ed.), *Animal Ecology*, New York: McGraw-Hill.
- Wei, J., M. Henrlock, D.J.A. Johansson and T. Sterner (2012), "The Fossil Endgame: Strategic Oil Price Discrimination and Carbon Taxation", *Journal of Environmental Economics and Policy*, **1**, 48–69.
- Wirl, F. (1994), "Pigouvian Taxation of Energy for Flow and Stock Externalities and Strategic, Non-competitive Energy Pricing", *Journal of Environmental Economics and Management*, **26**, 1–18.
- Wirl, F. (1995), "The Exploitation of Fossil Fuels under the Threat of Global Warming and Carbon Taxes: A Dynamic Game Approach", *Environmental and Resource Economics*, **5**, 333–52.



366 *Handbook of game theory and industrial organization: applications*

- Wirl, F. (2007), "Energy Prices and Carbon Taxes under Uncertainty about Global Warming", *Environmental and Resource Economics*, **36**, 313–40.
- Wirl, F. and E. Dockner (1995), "Leviathan Governments and Carbon Taxes: Costs and Potential Benefits", *European Economic Review*, **39**, 1215–36.
- Wirl, F., G. Feichtinger and P. Kort (2013), "Individual Firm and Market Dynamics of CSR Activities", *Journal of Economic Behavior and Organization*, **86**, 169–82.
- Xepapadeas, A. and A. de Zeeuw (1999), "Environmental Policy and Competitiveness: The Porter Hypothesis and the Composition of Capital", *Journal of Environmental Economics and Management*, **37**, 165–82.
- Yanase, A. (2007), "Dynamic Games of Environmental Policy in a Global Economy: Taxes versus Quotas", *Review of International Economics*, **15**, 592–611.
- Yanase, A. (2012), "Trade and Global Pollution in Dynamic Oligopoly with Corporate Environmentalism", *Review of International Economics*, **20**, 924–43.





14. Intellectual property

*Miguel González-Maestre**

1 INTRODUCTION

In this survey, we discuss the current literature on intellectual property, from a perspective that takes into account two main features of the evolution of modern economies:

First, the increasing level of complexity associated with the production and design of goods. This aspect is more relevant in some industries than in others. But it is clear that, in general, the development of new products is, increasingly, the result of assembling many parts, which, in turn, implies the combination of many previous ideas.

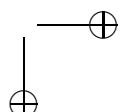
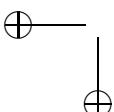
Second, the rapid development of the new technologies of information and communication has triggered the appearance of a new range of possibilities for both consumers and providers of cultural and technological creations. In particular, the expansion of the Internet, and other technological improvements, has substantially reduced the costs associated with the production and dissemination of new ideas and inventions (e.g., digital music, video-games, movies, word-processors).

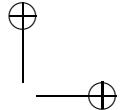
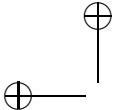
The growing complexity of the newly enhanced goods, mentioned in the first point above, has been associated with a remarkable debate on the role of patents as a way of stimulating innovation. In fact, some authors, like Boldrin and Levine (2008, 2012) have suggested the complete abolition of patents. However, in contrast with that view, a huge increase has been observed in the lobbying activity developed by incumbent owners of patents. Apparently, this lobbying activity has been rather successful, in view of the shift toward stronger U.S. patent rights in the early 1980s. In particular, after its creation in 1982, the Court of Appeals for the Federal Circuit (CAFC) promoted a broad interpretation of patent scope, which enhanced the broad exclusive rights of patent owners (Adelman, 1987; Merges, 1997).¹ In addition, the court reinforced patentees' rights by relaxing the conditions to grant preliminary injunctions to patentees during infringement suits (Lanjouw and Lerner, 1996). In Section 2 we will discuss this issue further.

Regarding the second point, the recent technological progress of the information and communication technologies should mean, in principle, good news from the social welfare perspective. However, the optimal regulation of intellectual property is crucial to ensuring this welfare-increasing effect. In fact, the above-mentioned technological improvements have been associated with a remarkable debate on how copyrights should be redesigned as a result of those rapid changes. In this respect, it seems that there is a substantial gap between the policy trends observed in most Western countries, which have been reflected in substantial extensions of copyright terms, and the opinion of many reputed economists against these

* I acknowledge financial support from the Spanish Ministry of Economy and Competitiveness, under projects ECO2014-53419-R and ECO2015-63679-P. The usual disclaimer applies.

¹ By comparing patterns of court decisions before and after the CAFC, Henry and Turner (2006) find a pro-patent shift associated with that institutional change.





policies.² In Section 3, we analyze this controversy in detail, in view of the theoretical and empirical evidence on copyrights.

The rest of the chapter is organized as follows: in Section 2 we discuss the literature related to the regulation of patents; Section 3 focuses on the literature dealing with copyrights; Section 4 considers the interplay between globalization and the political economy of intellectual property; and Section 5 gathers our main conclusions.

2 PATENTS

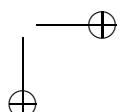
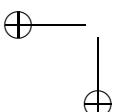
The debate on the optimal design of patent protection has been particularly intense in recent years (see *The Economist*, August 8, 2015). The classical view on the role of patents has been based on the trade-off between the underprovision and underutilization effects pointed out by Arrow (1962). According to this view, under a dynamic perspective, the underutilization of knowledge due to the monopoly distortion associated with patents is necessary to avoid the underprovision of knowledge associated with the lack of patent protection. One important implication of this traditional trade-off is that increasing the strength of patent protection should enhance innovative incentives. However, as we will discuss further, recent empirical studies tend to challenge this classical view.

The growing trend to reinforce patent rights in most Western countries has been subject to important criticisms from both the theoretical and the empirical perspectives. The empirical analysis developed by Lerner (2009) is, perhaps, one of the most robust pieces of evidence on the lack of a positive effect of patent protection on innovation. In fact, after a careful analysis of 177 of the most significant shifts in patent policy across 60 countries and 150 years, Lerner (2009) finds that the impact of patent protection-enhancing shifts on innovation is negative.

This issue is also discussed by Jaffee and Lerner (2004), who note the crucial institutional reform that took place in the U.S.A. in the early 1990s. Congress changed the structure of fees and financing of the U.S. Patent and Trademark Office (PTO), which became a kind of service agency whose costs of operation are covered by the patent applicants. According to Jaffee and Lerner (2004), the approval of new patents was much easier after this reform but, contrary to the conventional view, it did not have a clear positive effect on innovation. In fact, these authors point out that the alarming growth of legal uncertainty, triggered by the new regulation, is seriously threatening the incentives to innovate. Moreover, the strengthening of patent rights has enhanced firms' incentives to undertake what some authors call "strategic patenting." The idea is that, by increasing its patent portfolio, each firm tries to improve its strategic position in the market, in the face of future patent disputes. As pointed out by several authors, this gives rise to wasteful "patent portfolio races." We will discuss this issue in more detail further.

More specific empirical evidence has been documented, among others, by Bessen and Hunt (2007) in the case of the software industry in the U.S.A. Those authors investigate

² For example, the proposed extension of the copyright term in the European Union has been labeled as "a redistribution of income from living to dead artists" (Kretschmer, 2009, p. 2). Akerloff et al. (2002) and Liebowitz and Margolis (2005) provide different views on the optimality of the last extension of the copyright term in the U.S.A.



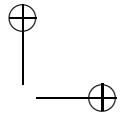
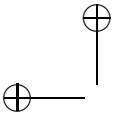
the natural experiment associated with increased patent protection for computer programs, particularly in the early 1990s. The authors find that, as a result of this shift in the regulation, the number of software patents has increased rapidly. Interestingly, according to the authors' analysis, this phenomenon has been dominated not by the software industry but by large manufacturing firms belonging to industries where strategic patenting seems to be crucial (e.g., computers, electronics, and instruments). The authors' findings are rather consistent with the theory that the pro-patent regulation changes have increased those firms' incentives to build strategic patent portfolios. This conclusion is reinforced by the fact that, according to the authors' empirical analysis, the large increase in software patenting cannot be explained by changes in R&D investments, or productivity growth. Therefore, their conclusions contradict the conventional view on the positive relationship between enhanced patenting and incentives to innovate. Instead, the hypothesis of strategic patent portfolios is consistent with the authors' results.

These empirical findings are consistent with some recent theoretical approaches explaining that the current patent system might reduce the incentives to innovate.

In particular, Bessen and Maskin (2009) have investigated this issue in a model with cumulative innovation and asymmetric information on future profits. The authors show that if innovation is sequential (each successive invention builds on its predecessors) and complementary (each potential researcher takes a different research line), then under reasonable conditions patent protection yields lower welfare than a system without such patent protection, based on competition and imitation. The basic intuition of this result relies on the idea that, under sequential and complementary innovation, free imitation helps to develop complementary ideas, which enhances the overall probabilities of future profits. The authors' assumption of asymmetric information on future profits is crucial (but reasonable) to ensure that free imitation can be superior to a patent system. Yet one more (reasonable) condition is needed for the free imitation system to work: the existence of some types of "frictions" affecting imitation. In Bessen and Maskin (2009) these frictions are reflected in the assumption of incomplete profit dissipation associated with competition from imitators.³ Obviously, this assumption is needed to ensure that innovators have an incentive to invest in R&D despite the competition faced under imitation.

In fact, it seems that the whole debate around the desirability of a patent system, versus a free imitation system, is implicitly based on the interplay between two types of "frictions": (i) the difficulties of optimal transactions in the patent system and (ii) the difficulties of imitation in a system without patents. Roughly speaking, the larger the degree of frictions (of both types), the more likely it is that the imitation system is socially superior to the patent system. In terms of a continuous trade-off, it could be said that the higher the level of frictions, the more permissive the system should be towards imitation. In the Appendix, we provide a formal illustration of this issue by developing a simple general model of intellectual property protection. In the rest of this section, we will further discuss the literature analyzing the regulation of patents, focusing on the interplay between these two frictions. To this end, Subsection 2.1 deals with the coordination failure, associated with patents, known as the

³ The frictions explaining this incomplete profit dissipation might be of different types, including learning costs, product heterogeneity (e.g., imitators' product is of lower quality), and imitators facing restricted short-run capacity or having higher marginal costs. In Section 3 we discuss a similar issue in the context of copyright.



tragedy of the anticommons, and in Subsection 2.2 we discuss the comparison between patents and imitation.

2.1 The Anticommons Effect

An important branch in the literature has emphasized criticisms of the patent system on the grounds that it yields a serious coordination failure known as the *tragedy of the anticommons*, a term coined by Heller (1998). In contrast with the *tragedy of the commons*, the inefficiency associated with the anticommons is due to the fact that “too many owners” of a non-rival good (e.g., the stock of patented ideas) prevent its optimal use. The anticommons problem was illustrated by Heller and Eisenberg (1998) in the case of biomedical research, but the idea has been extended to the analysis of other important industries, as we will discuss in our review of the empirical literature.

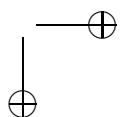
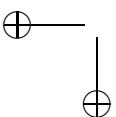
Murray and Stern (2007) investigate the empirical evidence of the anticommons effect in biotechnology, based on the concept of dual knowledge. The idea is that a single discovery may contribute to both scientific research (reflected in a paper) and useful commercial applications (which gives rise to a patent). The authors’ approach is to compare patterns of scientific citations to scientific articles that are part of patent–paper pairs, relative to citation patterns for articles that are not part of a patent–paper pair, but are similar along other dimensions. In the research, the authors exploit the fact that patents are granted with a substantial lag, after the publication of the associated papers. According to the anticommons hypothesis, the citation rate for a scientific publication should fall after a patent is formally granted (compared with the control group of articles that are not part of a patent–paper pair). In their empirical analysis of 169 patent–papers the authors find evidence of a modest anticommons effect: the citation rate after the patent grant declines by approximately 10 to 20 percent. Interestingly, very similar results have been obtained by Williams (2013) regarding the human genome research that has been associated with genes sequenced by the private firm Celera.

While those previous contributions focus on the specific biomedical field, Galasso and Schankerman (2014) compare the effects of patent protection on downstream innovation among different industries. Their strategy is based on the use of the patent invalidity decisions of the U.S. CAFC. According to the authors’ results, patent invalidation increases by 50 percent the citations to the focal patent on average. But this effect is rather heterogeneous across industries. The evidence supporting the idea that patents block downstream innovation is rather robust in computers, electronics and medical instruments, but not in drugs, chemicals or mechanical technologies. Interestingly, the authors find that this effect is basically driven by invalidation of patents owned by large firms, which increases the number of small firms that subsequently cite the focal patent.

Let us discuss some important aspects associated with the tragedy of anticommons: the *patent thicket*, the *patent trolls*, and the lobbying and rent-seeking effects of patents.

2.1.1 The patent thicket

The term “patent thicket” is used to mean how the technical complexity associated with the development of new products gives to rise to inefficient bargaining costs, as a result of the presence of multiple patents. As explained by Shapiro (2001), in many industries, like telecommunications and computers, a new innovator might face high transaction costs because of the need to bargain with many patent owners. This is known as the fragmentation of



patents.⁴ As pointed out by Stiglitz (2008), the patent disputes around the early developments of the airplane in the U.S.A. are one of the most remarkable examples of the dramatically harmful effects of patent thickets on the incentives to innovate.⁵

The theoretical analysis of this issue has been recently considered, among others, by Llanes and Trento (2012). Those authors show that under multiple sequential innovations the patent system might be the wrong way of regulating innovation and, in consequence eliminating patent protection would improve welfare in complex industries such as electronics, software and hardware.

Recent empirical contributions have investigated the *patent thicket* effect, from different perspectives. Galasso and Schankerman (2010), analyze the relationship between the fragmentation of patents and the duration of patent disputes in five US industries. The authors focus on the analysis of two opposite effects associated with the increasing fragmentation of patents: the *thicket effect* and the *negotiation value effect*. On the one hand, fragmentation tends to increase the total negotiation time because of the need to negotiate with many patentees (the *ticket effect*), but, on the other hand, fragmentation reduces the value at stake in each negotiation, which reduces the negotiation time per dispute. In their empirical estimations, the authors find out that, under the current U.S. regime for patent litigation, which is centralized at the CAFC, the *ticket effect* dominates the *negotiation value effect* for all the analyzed industries.

One important aspect associated with the *patent thicket*, is related to each firm's incentive to increase its patent portfolio for strategic purposes. In particular, Hall and Ziedonis (2001) have found empirical evidence, in the semiconductor industry, suggesting that the 1980s' strengthening of U.S. patent rights triggered "patent portfolio races" among capital-intensive firms. In a related work, Hall and Ziedonis (2007) have focused on the link between this shift in the U.S. patent regulation and the level of litigation in the semiconductor industry. They find robust evidence that the strengthened patent rights have produced an escalation in litigation against semiconductor firms brought by outside patent owners. The authors conclude that this patent holders' behavior is reflecting an "ex post holdup" strategy. In an extensive analysis, related to this issue, Lanjouw and Lerner (1996) have examined a sample of 252 patent suits. They find that their data is consistent with the hypothesis that preliminary injunctive relief (which was enhanced after the U.S. regulatory reform in the 1980s) is a predatory tool that favors large firms over small firms in patent cases.

One of the perverse effects of the patent thickets is related to firms' incentives to build strategic "patent portfolios." The idea is that each firm has an incentive for holding patents in order to make credible the threat of bringing a potential suitor to court in the case of litigation. However, this strategic effect gives rise to wasteful patenting and litigation costs. A theoretical model on this issue is analyzed by Bessen (2003).

As pointed out by Boldrin and Levine (2012, p. 74) the situation is similar to that associated with the "arms race" during the Cold War. Jell and Henkel (2010) provide recent empirical evidence of this effect in the newspaper printing machines oligopoly.

⁴ In Shapiro's words (p. 120), the *patent thicket* is defined as "a dense web of overlapping intellectual property rights that a company must hack its way through in order to actually commercialize new technology."

⁵ As explained by Stiglitz (2008), the Wright brothers obtained some key patents needed to develop an airplane, but other key patents were granted to another innovator, named Glenn Curtiss. In the face of this patent thicket, potential developers of the airplane anticipated too much expected bargaining and litigation costs and, as a result, the airplane was not developed in the U.S.A. until World War I, when the U.S. government seized the patent.

2.1.2 The patent trolls

The term “patent troll” is a pejorative expression for the owner of a patent who does not manufacture products. The related term “non-practicing entity” (NPE) basically means the same but without the negative connotation associated with patent trolls. In principle, NPEs could play a useful role for protecting innovators’ rights. However, some recent contributions have emphasized that rather than helping innovators, NPEs tend to act as opportunistic rent-seekers, discouraging innovation. This negative effect explains the pejorative term “patent trolls”. In particular, Bessen, Meurer and Ford (2011) have undertaken interesting empirical work on this issue, using an extensive database of NPE lawsuits in the U.S.A. for the period between 1990 and 2010. They conclude that the lawsuits increase NPEs’ incentives to acquire vague, overreaching patents, but they do not increase incentives for real innovation.

In an extensive study of patents and lawsuits between NPEs for the period 2001 and 2011, Cohen, Gurun and Kominers (2014) show that the probability of being sued by NPEs increases dramatically with firms’ cash balance, and also increases if a firm employs fewer lawyers. In addition, the authors find that the large firms are the worst offenders and that after a firm is sued by an NPE it substantially reduces its innovation activity substantially. Moreover, in a related work, Bessen and Meurer (2014) estimate that when NPEs win infringement suits only 5 percent of the awarded damages is paid back to innovators. The rest goes to lawyers and NPEs that do not innovate.

In comparison with the previously discussed “patent portfolio races” among practicing entities (PEs), the strategic use of patents works differently in the case of patent trolls. In the case of patent trolls the NPE has a stronger strategic advantage over the PEs because its losses are zero in the case of litigation (it cannot be countersued). As explained by Boldrin and Levine (2012), this feature of patent trolls tends to reduce innovation incentives and welfare with respect to a system without patents. The idea is that, under a patent system where taking out a patent is very cheap, some firms might prefer to litigate with real innovators instead of producing a marketable product. Those authors mention the illustrative case of Microsoft litigating with Google in the smartphone market. Instead of developing a new marketable product, Microsoft just tried to get a share of Google’s revenue.

2.1.3 Lobbying and rent-seeking

One of the justifications of patents is associated with the conventional view that the patent system enhances the innovators’ incentives for the disclosure of their inventions, rather than keeping them secret. In other words, public rent-seeking (patents) are substitutes for private rent-seeking (secrecy). However, Boldrin and Levine (2004) have shown that this is not necessarily the case. Instead, the presence of a patent system, the authors argue, can even increase the inventors’ incentives for private rent-seeking in the form of patent litigation and lobbying. Apart from the increased monopoly distortion, those activities involve additional wasteful costs for society.⁶

On the empirical side, Bessen and Meurer (2009) have undertaken an extensive investigation into how the increased patent protection in the early 1990s (particularly in the software industry) has been associated with a substantial increase in lobbying and rent-seeking activities. According to those authors, rather than enhancing innovation, these extensions of patent protection are having harmful effects on the incentives to innovate. One of those

⁶ See the interesting discussion by Stiglitz (2008) on this issue.

harmful effects arises from what the authors call the “notice failure”: the inability to provide predictable property rights. This effect results in a higher degree of legal uncertainty that undermines the economic utility of patents by reducing innovation incentives and increasing litigation costs. The authors conclude that, in most industries, the current patent system triggers excessive uncertainty and litigation costs that outweigh the positive incentives on innovation. A similar conclusion is reached, among others, by Burk and Lemley (2009) and Jaffee and Lerner (2004).

As noted by Boldrin and Levine (2012), the lack of research about the political economy of patents is surprising. In one of the few contributions on this issue, Landes and Posner (2004) recognize the huge increase in the lobbying activity by advocates of patents, during the last 30 years, in contrast with the relatively much smaller effort by opposers of patents. These authors address the following question on regulatory changes in the recent decades: why have some specific regulations (e.g., those affecting intellectual property) increased dramatically in the middle of a general antiregulatory trend? They suggest that the creation of the U.S. CAFC as a specialized court, with the “mission” of promoting technological progress by enlarging patent rights, might have been the consequence of the lobbying activity of the American patent bar.⁷ This effect is reinforced by the asymmetric incentives for lobbying in the context of patents. The idea is that the per-capita increase in the monopolistic rents of patent holders is much higher than the individual deadweight loss of each individual consumer. This rent-seeking argument is consistent with the empirical analysis by Landes and Posner (2003a), which confirms that the creation of the CAFC explains the huge increase in the number of patents applied for and granted but it had no positive effect on the rate of technological progress.

In a recent empirical work, Kesan and Gallo (2009) find robust econometric evidence demonstrating that the passage of the bill reforms in the patent system in the U.S. House of Representatives was strongly correlated with the resources that the primary stakeholders provided to each of the congressional representatives. The authors conclude that their analysis shows that the U.S. Congress does not have an independent point of view from the stakeholders in the patent system.

2.2 Imitation versus Patents: The Role of Frictions

Boldrin and Levine (2008, 2012) provide extensive discussion showing that first, incentives are higher under imitation than under a patent system. This is due to the role played by the involved frictions. (a) Under the patent system, the empirical evidence shows that the innovation incentives are seriously reduced as a result of the transaction costs associated with the various forms of the “tragedy of the anticommons.” (b) Because of a number of frictions associated with imitation, patents are not necessary to ensure innovation incentives, as has been shown in many empirical examples. In particular, those authors mention the crucial role of costly imitation, short-run capacity restrictions, and the usually long period of time needed to learn the new technology, as factors ensuring that innovative firms can recover their research investment even in the absence of patents. The authors mention several remarkable historical examples of successful innovative industries without patents: the rapid

⁷ As pointed out by Boldrin and Levine (2012) patent lawyers have an obvious incentive to see that more patents are issued (according to Quinn, 2011, legal fees for filing a patent run upwards of \$7,000). One of the most remarkable examples of the successful lobbying pressure of patent lawyers is the 1994 Tektronix decision expanding the scope of patents to software.

development of the software inventions before the 1980s (when software was not patentable), the irrigation systems introduced by competitive farmers in the driest of the Spanish provinces (Almería), the creative franchising techniques introduced in the sweater industry in Treviso (Italy), and many others. In a related work, Boldrin et al. (2011) find robust econometric evidence, using an extensive micro data set, showing that there is, in general, no statistically significant correlation between measures of productivity and patenting activity. The authors argue that this result suggests the use of patents either as a defensive or a rent-seeking tool.⁸ In addition, the authors find a positive relationship between innovation (measured by patents or patent citations) and competition (measured as the inverse of profitability).⁹

Second, lobbying for imposing an enhanced patent system arises when the industry is mature and the new competitors have developed a large capacity. Basically, the old incumbents try to exploit their position as patentees to extract the rents associated with the increased capacity of new competitors. In fact, the empirical literature provides strong evidence, in the evolution of many industries, showing the prevalence of a competitive innovation system, based on imitation, at the early stages of the market. It is at the later stages of the product cycle when the oldest innovators try to patent old ideas to extract economic rents from the rest of the firms. Moreover, as previously discussed, in many cases, the old innovators do not even compete with the latest innovators. Instead, they just use their patent trolls for purely rent-seeking purposes.

In view of those arguments, some authors have proposed that, rather than reforming the current patent system, what is needed is the complete abolition of patents. In particular, Boldrin and Levine (2012) argue that even if we think that the first-best regulatory system involves some degree of patent protection, the political economy about patents, which has been previously discussed, suggests that, as a result of rent-seeking and lobbying incentives, the first-best would be manipulated by some players. In consequence, the second-best policy should imply a credible commitment for complete abolition (e.g., by means of a constitutional change). However, even supporters of this drastic solution should recognize the difficulties for implementing this type of policy, at least in the short run. Perhaps a more realistic approach could be based on the differential effects of patents across industries, as documented by the empirical evidence. Therefore, an alternative policy could be based on the design of industry-specific patent policies, depending on the empirical connection between patent protection and innovation incentives.¹⁰

3 COPYRIGHTS

As in the case of patents, the debate on copyrights has intensified in recent decades. The rapid development of the new technologies of information and communication, and the increased globalization of markets has been associated with different reactions from the involved parties. On the one hand, copyright holders of artistic creations have increased their pressure to obtain,

⁸ See the theoretical analysis on this issue by Boldrin and Levine (2004).

⁹ The authors use both an original micro data set and an enriched version of U.K. micro data set used by Aghion et al. (2005) in a previous work yielding an inverted “U” relationship between mark-ups and innovation.

¹⁰ Nevertheless, those potential changes will always be subject to the political pressure by interested lobbying groups, as has been shown in the paper by Kesan and Gallo (2009) already mentioned.

rather successfully, substantial extensions of the copyright term.¹¹ On the other hand, many academic researchers have criticized these types of policies on the grounds that enhanced copyrights could harm both social welfare and the incentives for artistic creation.

Similarly to what happens in the case of patent regulation, the debate on copyrights can be interpreted in terms of different views on the role played by different frictions. In particular, supporters of enhanced copyrights argue that the recent advances in information and communication technologies substantially reduce the frictions associated with copying original creations. This is reflected in the unauthorized copying, or piracy, affecting the music industry and other relevant activities like video-games, movies, or literary creations. Once unauthorized copying becomes less costly and easier, original creators find it more difficult to cover their opportunity costs and, according to conventional wisdom, creative incentives fall. In a classical paper on this issue, Novos and Waldman (1984), analyze this effect in a model where unauthorized copying reduces a monopolist's incentives to invest in quality.

However, critics with this view remark that the reduction of other types of frictions, associated with technological progress, substantially favor original creators. In the case of the music industry, the possibility of reaching large audiences has been enhanced dramatically with the emergence of radio, recorded music, television, and, more recently, the Internet. Because this aspect of technological progress tends to increase artists' earnings, it suggests that copyrights should be weakened, rather than reinforced, in order to reduce the monopoly distortion associated with copyrights. In particular, one interesting issue that reinforces the arguments in favor of a lenient policy towards unauthorized copying relates to the informational role of copying. The idea is that, because music and other similar creations are experience goods, many Internet users take advantage of modern technological devices (e.g., file sharing) to obtain information on how a particular product (a song, a movie or a video game) fits their preferences. As a result of this sampling effect, the consumer's willingness to pay (and hence producer profits) might be enhanced (see Belleflamme and Peitz, 2012, for a detailed discussion of different theoretical models dealing with this issue). Obviously, this effect suggests an extra argument for a permissive policy about copying.

In addition, some aspects of this technological progress have favored a huge increase in what Rosen (1981) called the "superstar effect." The idea is that because of the increased top artists' ability to reach large audiences, earnings differences between artists become much more than proportional to quality differences. As a result, substantial monopoly rents are granted to top artists, and those rents increase when copyright protection is enhanced. In turn, some authors have emphasized, recently, that this "superstar effect" might be substantially harmful, in the long run, for both artistic creation and welfare. This issue has been analyzed theoretically by Alcalá and González-Maestre (2010, 2012), and it will be discussed in more detail further.

Based on these preliminary ideas, in the rest of this section we will discuss the literature on copyrights focusing on the following issues:

- (i) the role of derivative works;
- (ii) the copyright term;

¹¹ Particularly remarkable is the recent extension, in the U.S.A., of the copyright term from 50 to 70 years after the death of the creator.

- (iii) the effects of piracy;
- (iv) copying levies;
- (v) the role of collecting societies and market power in creative industries;
- (vi) the role of new business models.

3.1 The Role of Derivative Works

As previously explained, some aspects associated with the regulation of copyrights suggests that copyrights are needed to maintain the incentives for artistic creation, but others suggest that increasing copyright protection might be harmful for welfare and creative incentives. Next we will discuss the literature dealing with some of those aspects. The paper by Landes and Posner (1989) is one of the first theoretical contributions pointing out that excessively strong intellectual property rights may in fact hinder the development of new ideas that are based on previous ones. According to those authors there is a trade-off between easing the access of new creators to previous knowledge (which enhances the development of new creations) and providing incentives for the creation of that previous knowledge. As a result of this trade-off, the authors show that the level of copyright protection that maximizes social welfare is lower than the level that maximizes artistic creation.

Other aspects associated with derivative works have been developed by recent contributions. Motivated by the massive copyright infringement around the *anime* and *manga* characters in Japan, Arai and Kinukawa (2014) analyze the effects of the unauthorized derivative works based on those original characters. To this end, the authors develop a model that incorporates positive and negative externalities:

First, the positive externalities considered in their model are based on the previous contribution by Mehra (2002), who emphasizes that the markets for this type of derived works (labeled *doujinshi* in Japan) help to promote new talented creators, enhances the sales of the original works, and generates new ideas that can be incorporated in commercial *anime* and *manga*. These derivative works are mainly developed by amateurs who are, at the same time, consumers of the original *anime* and *manga*. This means that the markets for these derivative works are a form of user innovation, which is more commonly observed in technological markets.¹²

Second, the negative externalities appearing in this model are based on the potential misuse of the original creations, as indicated by Liebowitz and Margolis (2005), and on the congestion effects that might arise from the proliferation of the characters of the original work, as pointed out by Landes and Posner (2003b).

After analyzing the combined effects of both positive and negative effects, Arai and Kinukawa (2014) show that ignoring copyright infringement by a derivative creator can be optimal not only from the social welfare perspective but also for the copyright holder of the original characters, provided that the market size is not too large. The authors argue that this result could help to explain why the copyright holders of commercial *anime* and *manga* have been ignoring copyright infringement by *doujinshi* creators. However, their model predicts that as the market for derived work expands, then copyright owners may not allow unauthorized use by derivative creators in the future, even if ignoring the unauthorized

¹² This effect has been documented, among others, by Henkel and Von Hippel (2005).

use is socially optimal. The authors suggest that a remedy for the discrepancy between the right holder's profit and social welfare can be a lenient application of the fair use doctrine.¹³

3.2 The Copyright Term

Alcalá and González-Maestre (2012) have analyzed the relationship between the extension of copyright term and artistic creation in a model that takes into account three important aspects of many artistic markets: (i) the superstar effect working in those markets (Rosen, 1981); (ii) the crucial role of the promotion and marketing expenditures in determining market shares; and (iii) artistic talent – which is a key ingredient in producing artistic ideas – is sorted out and developed through artistic careers that most often end in failure.

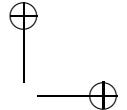
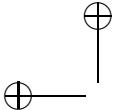
On the basis of those ingredients, the authors develop a model of overlapping generations of artists. Their main finding is that increasing copyright term might reduce both the level and average quality of artistic creation. This result holds if the “superstar effect” is strong enough (which is reasonable in this type of markets). Intuitively, stronger copyrights enhance superstars' incentives to invest in marketing and promotional expenditures, which increases their market share and works as a barrier to entry for potential new artists. In consequence the dynamic process of sorting out the potential future talent is weakened by enhanced copyrights and the long-run level of high-quality artistic creation is reduced. In fact, it is shown that, from the social welfare perspective, the stronger the superstar effect, the shorter the optimal copyright term should be.

Interestingly, this result contrasts with the observed trend in recent decades to enlarge the copyright term in most Western countries, but it is consistent with the political economy discussed in the previous section. As in the case of patents, the higher the level of monopoly rents associated with enhanced copyrights, the higher the lobbying and rent-seeking incentives to get those rents.

Alcalá and González-Maestre (2012) emphasize that the modern development of technological advances enhances the creators' ability to reach larger audiences (e.g., radio, TV, Internet). In addition, those advances tend to increase the effectiveness of promotion and marketing expenditures of top artists, which reinforces the stardom effect. The crucial role of those expenditures has been widely documented in the empirical literature. As reported by Peitz and Waelbroeck (2005), in the music industry marketing and promotion are often the main cost of making and selling a CD.

The empirical evidence about the growing superstar effect in recent decades has been documented in several contributions. In particular, Krueger (2005) reports that in Rock and Roll music the top 1 percent of artists obtained 26 percent of concert revenue in 1982. In 2003, this proportion increased to 56 percent. Kretschmer and Hardwick (2007) report data from the distribution of payments in 1994 by the U.K. Performing Rights Society. According to the data, the top 9.3 percent of writers earned 81.07 percent. In Spain, top 1.5 percent of beneficiaries of the main collecting society in the country, SGAE, obtain 75 percent of total revenues (AEVAL, 2008). Rothenbuhler and Dimmick (1982), Crain and Tollison (2002), and Pitt (2010), among others, provide additional evidence on the market concentration in

¹³ Interestingly, the relatively weak legal regime in Japan seems to have played a positive role in the development of the *doujinshi* markets. As pointed out by Mehra (2002, p. 155) this legal weakness “has by chance solved a collective action problem and prevented the interests of a few copyright holders from inhibiting the growth and development of the industry as a whole.”



the music industry. Similarly, in the motion-picture industry, Walls (2005) found that a small proportion of successful films earn the majority of box-office revenue.

Apart from promotional and marketing investments, there are other important aspects of artistic markets that reinforce the stardom effect. In particular, Berlin, Bernard and Fürst (2015) have undertaken an experimental research showing that word-of-mouth communication among consumers tends to lower diversity in artistic markets. This result is similar to the one obtained by Salganik, Dodds and Watts (2006) in a controlled online experiment that shows that observing other individuals' behavior increases the skewness of the distribution of the demand for cultural goods. Those experimental results are consistent with the theoretical analysis by Adler (1985, 2006). According to this author, because the consumption of cultural goods requires information and knowledge, demand tends to concentrate on the most popular artists, even if quality differentials are negligible.

As pointed out by Handke (2012), according to the empirical evidence, it seems that increasing copyright strength does not have a substantial effect on creative incentives. In particular, Khan (2004) finds that the number of full-time authors did not increase significantly after the U.S. International Copyright Act of 1891. In the music industry, Scherer (2008) investigates copyright extensions in Europe between 1709 and 1850 and finds that market entry by composers did not change significantly.¹⁴ Landes and Posner (2003a) find no substantial effects of the term extensions in 1962 and 1998 on the number of optional U.S. copyright registrations.

The empirical relationship between copyright term and artistic production in the movie industry has been investigated, among others, by Png and Wang (2009). Using a panel of 23 OECD countries, among which 19 extended copyright term at various times between 1991 and 2005,¹⁵ the authors found no statistically robust evidence that copyright term extension was associated with higher movie production. This empirical result suggests that rather than enhancing creativity, copyright term extension has just increased stars' monopoly rents in the movie industry. Interestingly, this conclusion is consistent with the empirical work by Chisholm (2004) indicating that stars obtain substantial economic rents in the motion-picture industry.

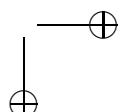
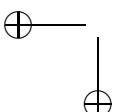
3.3 The Effects of Piracy

Some recent theoretical literature suggests that unauthorized copying of original artistic creations might have both positive and negative effects on the incentives for artistic creation.¹⁶ The interplay between these effects depends on the role played by different types of frictions. In particular, Novos and Waldman (1984) assume, in the context of a monopoly innovator, that the quality of the pirated good is the same as the original one, and that all the consumers have the same valuation of the good. The only friction in their model is determined by the assumption that copying is costly. Under some conditions, the authors show that increasing

¹⁴ Interestingly, the historical analysis by Tschmuck (2002) shows that a great deal of classical European artistic creation was developed without the intervention of copyright protection.

¹⁵ Beginning in 1993, various European countries extended the term of copyright from 50 to 70 years. The same policy was followed by the U.S.A. in 1998.

¹⁶ As noted by Varian (2005), these two variables are endogenously determined: the same technological advances that lower the costs of unauthorized copying are also helping to reduce the fixed cost of creating and distributing new content, which would increase the production of new creative works.



the enforcement of copyrights (which increases the consumers' cost of copying) increases the firms' incentives to increase quality, which implies higher welfare levels. However, other authors have shown, more recently, that if other frictions are allowed in the model then the welfare conclusions are substantially different. In particular, Bae and Choi (2006) assume, under the same basic monopoly model, that the quality of the pirated copy is lower than the original good, and that consumers have heterogeneous valuations of the unauthorized copy. They conclude that, in contrast with the model by Novos and Waldman (1984), increasing the enforcement of copyrights may reduce the long-run level of welfare.

As previously indicated, part of the theoretical works dealing with piracy have focused on the informational or sampling effect associated with different forms of unauthorized copying (e.g., file downloading or file sharing). In fact, because of the potential increase in the willingness to pay for the original work, this effect might even be profitable for the sellers of original copies of artistic creations. If this effect is strong enough then the net effect of piracy could be positive from the perspective of creative incentives. The idea is that sampling tends to reduce one important source of frictions in the market: the lack of information. There are several papers exploring different aspects of this issue.¹⁷ In particular, Ahn and Yoon (2009) consider a monopoly model where sampling increases the consumers' willingness to pay for the original good. The authors show that if the sampling effect is enhanced (due to the impact of digitalization) then welfare might increase in the long run. Peitz and Waelbroeck (2006) have extended this analysis in the context of a multiproduct environment. These authors show that if the number of products and the degree of product differentiation are sufficiently large, the firm can profit from the informational role of digital copies. Takeyama (2003) analyzes the frictions due to adverse selection problems. This author shows that copying might solve the underprovision of digital products appearing when the consumer has less information about the product quality than the firm.

There are other ways that unauthorized copying might enhance the incentives for artistic creation. In particular, Kim (2007) has emphasized the strategic use of the copyright system as a barrier to entry in artistic markets. Kim's work is motivated by the observation that while most of the stars were against digital copying,¹⁸ relatively unknown artists endorsed Napster for its promotional role for new artists. In particular, according to an extensive online survey among musicians in U.S.A., 83 percent of the musicians "offer free samples online and notable numbers report benefits from that such as higher CD sales, larger concert attendance, and more radio play."¹⁹ Based on those observations, Kim (2007) analyzes a model in which the incumbent artists (superstars) influence the level of copyright protection (e.g., by lobbying) for strategic purposes. One important ingredient in the model is that besides the primary source of revenue for artists (record sales) there are complementary sources (concert revenues and radio broadcasting). The authors show that if the complementary sources of revenues are sufficiently important then strong copyright protection discourages entry in the sense that the probability of entry is reduced. According to the authors, an implication for policy-makers is

¹⁷ Banerjee (2003, 2006) and Martínez-Sánchez (2010) analyze the strategic interactions among firms, pirates and government, in the context of commercial piracy. Belleflamme and Peitz (2012) provide an extensive and detailed discussion of the literature dealing with digital piracy.

¹⁸ Kim (2007) mentions the anti-Napster campaign launched by the group called Artists Against Piracy, which was organized by more than 70 well-established artists in the U.S.A. A similar coalition, organized in Britain was led by Paul McCartney and Elton John, who are the richest and the second richest rock stars.

¹⁹ Pew Research Center. Report 12/5/2004. Available at <http://www.pewinternet.org/2004/12/05/artists-musicians-and-the-internet/>.

that they should be more cautious about overprotecting the established stars to the detriment of new entrants.

The potential entry-deterring effect of too strong protection against file sharing and unauthorized copying has also been analyzed by Alcalá and González-Maestre (2010) in a dynamic model of overlapping generations of artists. These authors show that the emergence of piracy tends to reduce the stardom effect in a similar way to the decrease in the copyright term considered in Alcalá and González-Maestre (2012). In the model of piracy developed by these authors, unauthorized copying mainly harms superstars' rents, which reduces top stars' incentives to invest in promotional and marketing costs. As a consequence, the entry barriers associated with those investments are also reduced and, as the authors show, the long-run artistic creation and welfare increase, under reasonable conditions, if copying is allowed.

While these contributions emphasize the conflict of interest between superstars and young or modest artists, Gayer and Shy (2006) have analyzed the conflict of interest between artists and publishers, regarding piracy. This conflict arises because, as we have already noticed, artists also earn their profit from other market activities such as giving live performances, in addition to their share of profits from sales of their copyrighted recordings via the publishers. The authors identify the conditions under which publishers of recorded media may lose from piracy, whereas artists may gain from piracy. The authors conclude that massive antipiracy campaigns, and the large number of ongoing and pending law suits observed in many countries (which are mainly undertaken by publishers and recording firms), may eventually hurt artists. Because the contractual relationships between artists and publishers usually means that the artists' share in copyright revenues is a rather small proportion of their total revenues (see Caves, 2000; Towse, 2001; and Connolly and Krueger, 2006), it seems that this conflict of interest is very relevant. In turn, this argument reinforces the idea that the rigorous antipiracy policies that are strongly influenced by lobbying pressure from publishers are very likely to be socially harmful.

On the empirical evidence,²⁰ Waldfogel (2012) reports that legal revenue from recorded music declined from \$37 billion in 1999 to \$25 billion in 2007 as a result of the appearance and rapid growth of the Napster file-sharing service. However, this dramatic reduction in revenues does not seem to have reduced creative incentives. Oberholzer-Gee and Strumpf (2009) report that the number of albums created has increased dramatically in the period in which file sharing has become widespread. In 2000, 35,516 albums were released. Seven years later, 79,695 albums (including 25,159 digital albums) were published (Nielsen SoundScan, 2008). According to Waldfogel (2012) these findings suggest that the short-run benefits obtained by consumers, as a result of unpaid consumption, have not been counterbalanced by a reduction in available creative works.

The long-run effects of unauthorized copying have been investigated also by Handke (2012). This author provides empirical evidence that digital copying has not reduced the supply of new copyrighted sound recording in Germany, in the period between 1999 and 2006. In addition, the paper also presents evidence that the amount of time listening to sound recordings has not fallen over this period, suggesting no strong decline in the quality of new work. These empirical findings are consistent with the theoretical contributions we have discussed above. The author suggests that these results might be due, at least partially, to the combination of two effects. First, advances in information technology could be associated

²⁰ See Gomes, Cerqueira and Almeida (2015) for a recent survey on the empirical aspects of software piracy.

with lower fixed costs per product variant, which could help explain a greater number of new artistic products²¹ despite falling revenues.²² Second, creators and rights holders might have adapted to the impact of digital copying by increasing the role of related markets, such as live music, music licensing or merchandising.

Recent empirical literature reinforces this view. In particular, Mortimer, Nosko and Sorensen (2012) examine the impact of file sharing on sales of recorded music and on the demand and revenues for live concert performances. Using concert data on over 200,000 concerts between 1995 and 2004 in the U.S.A., they find robust evidence suggesting that while file sharing reduced album sales, it simultaneously increased demand and revenues for concerts.²³ Moreover, they also show that these two effects are not symmetric among artists. The increase in the demand for concerts is most pronounced for small artists, which suggests that file sharing boosts awareness of such artists. In contrast, the decrease in album sales is negligible in the case of smaller lower-ranked artists but very high in the case of superstars or top-ranked artists. Due to these asymmetric effects, the authors conclude that their results imply that after the appearance of Napster top-ranked artists lose market share in both markets (live concerts and recorded music) in favor of smaller artists. In a similar spirit, Gopal, Bhattacharjee and Sanders (2006) analyze a model of piracy in which sharing technologies erode the superstar phenomenon widely prevalent in the music business. Those authors show that top artists lose from file sharing, but less popular artists gain from the extra exposure and lower distribution costs associated to Internet sampling. The overall effect of file sharing is shown to be welfare enhancing in the long run. The authors conclude their research with an extensive empirical investigation, based on surveys and Billboard ranking charts, showing that their theoretical analysis is validated by their empirical results. More recently, Albinsson (2013) has used annual reports of the STIM (Swedish Performing Rights Society) finding that the digital technology shift has resulted in illegal downloading that explains a decrease in total revenues for composers from record sales. However, the author shows that there has been a simultaneous growth in income from other sources, which compensates for the loss from record royalties. Moreover, the author also finds that a very small group of composers receives a very large share of the copyright revenues, which implies that only top artists lose significantly from file sharing.

Interestingly, these empirical findings are consistent with the previously discussed theoretical works by Kim (2007) and Alcalá and González-Maestre (2010) suggesting that a permissive policy towards the use of copying devices might enhance artistic creation (and welfare) through its effect in reducing the entry barriers faced by new potential artists.²⁴

²¹ This effect has been extensively analyzed by Anderson (2006) and has given rise to what this author calls “the long tail.” As a result of the development of the new information technologies, a new “long tail” of small artistic products has appeared, reducing the market share of traditional hits or blockbusters. Anderson (p. 54) illustrates this effect in the music industry, reporting that “the number of new albums released grew a phenomenal 36 per cent in 2005, to 60,000 titles (up from 44,000 in 2004, largely due to the ease with which artists can now record and release their own music.” See the theoretical analysis of this issue by Bar-Isaac, Caruana and Cuñat (2012).

²² Moreover, this author suggests that the share of creators in total record industry revenues could have increased due to greater competition among intermediaries.

²³ Using the 2006/2007 wave of Spain’s *Survey on Habits and Cultural Practices*, Montoro-Pons and Cuadrado-García (2011) estimate a bivariate probit model for attendance at live concerts and the purchase of prerecorded music. They find evidence of demand complementarities, with a direct causal link from prerecorded music to live attendance.

²⁴ As indicated by Handke (2012), the idea that unauthorized copying may increase the contestability of the market by hurting well-established incumbents is widely supported by the empirical evidence. Apart from the previously mentioned works, see Handke (2006, 2010) and Bhattacharjee et al. (2007).

3.4 Copying Levies

The current system of copying levies, growingly implemented in most Western countries as a way of compensating the creators for the private copying of original artistic productions,²⁵ has been subject to strong criticism by many authors.

Legros and Ginsburgh (2013) argue that when levies are not targeted, such as levies on general-purpose computers, they may reduce the demand for hardware without substituting the tendency to copy with demand for the original content. Therefore, copying levies create a substantial distortionary effect in other industries. In an extensive research on the use of copy levies in the E.U., Ferreira (2010) reveals that copyright levies cause an economic loss of at least 51.2 cents for each euro collected.²⁶

Alcalá and González-Maestre (2010) find that copyright levies may reduce both the short-run number of young artists and the long-run supply of high-quality artists. The argument behind this result is based on the “superstar” effect associated with increased copyright protection. Because copy levy revenues are concentrated on top artists they have an incentive to increase promotion and marketing investments in order to increase their market share. As a result of this effect, entry by new young artists is reduced, which implies lower artistic production in the long run. In contrast, the authors argue that artistic creation can be stimulated more efficiently by allocating levy revenues among artists according to a non-linear scheme that favors young artists.

Kim (2013) has explored the political economy of copying levies in a model where the collecting societies use a discretionary budget for lobbying activities.²⁷ In the context of a closed economy, the author shows that if the lobbying pressure of collecting societies is high enough then the equilibrium levy rate is above the socially optimal. The author extends the model to analyze the effects of international harmonization of copy levies. The model is calibrated using European data showing that harmonization would increase aggregate social welfare. However, the author also finds that, because of lobbying pressures, policy-makers are worse off in some countries as a result of harmonization (although consumers are better off). The author suggests that this may explain why the efforts to harmonize the levy rates by the European Commission have been unsuccessful. The author points out that the main effect behind this result is driven by the presence of large countries with high lobbying pressure from collecting societies. Interestingly, it seems that this might be the case with France. In contrast, other large countries with lower lobbying pressure (e.g., the U.K.) have rather small or even non-existing copying levies.²⁸

3.5 The Role of Collecting Societies and Market Power in Creative Industries

The current regulation of the collecting societies, which are engaged in collecting the copyright royalties and distributing the revenues among copyright holders, has been subject

²⁵ According to Kretschmer (2011), across Europe, there are great variations in the products subject to copyright levies. There are levies on blank media in 22 EU countries, on MP3 players in 18 countries, on printers in 12 countries, and on personal computers in four countries.

²⁶ The extensive report by Kretschmer (2011) analyzes in detail the administrative and economic inefficiencies associated with copying levies.

²⁷ See Huang and Png (2010) for empirical evidence of lobbying activities around copying levies.

²⁸ Kretschmer (2011) reports that levy density in 2009, measured by revenues raised per capita of the population, ranges from €2.6 in France to €0 in non-levy countries, such as the U.K. and Ireland.

to strong criticisms in recent years. In particular, Katz (2005) has emphasized that the usual arguments tending to justify the promotion of these collecting societies are based on the idea that they are natural monopolies. By focusing on the role of performing rights organizations (PROs) in the music industry, this author discusses critically the arguments supporting the role of PROs as natural monopolies. In contrast with the conventional view on this issue, Katz (2005) argues that many of the underlying cost efficiencies that are attributed to PROs are far from clear and/or in many cases could be equally achieved under less restrictive arrangements than those associated with the regulation of those entities. According to this author, the monopoly distortions associated with PROs can be overcome by enhancing competition among several collecting societies. This idea is related to the previous contribution by Besen, Kirby and Salop (1992). Instead of a government regulation, these authors propose an alternative competitive licensing system in which the members of a collective compete in issuing licensing but cooperate in the administration of their respective rights.

More recently, Ferreira (2010) summarizes several critical reports by official institutions of Member States of the E.U. regarding the regulation of collecting societies. In essence those reports are strongly critical with the anticompetitive practices of the collecting societies operating in many European countries, including abuse of dominance position, lack of accountability, and complex and large management costs, among other criticisms. To illustrate the potential social benefits associated with a more competitive framework, Ferreira mention the report by the French Committee on the Immaterial Economy, in 2006. This report takes the example of Japan where in 2001 the government ended the monopoly of the Japanese collecting society JASRAC. After this pro-competitive shift, four competitors entered the market, which remarkably reduced the fees charged by JASRAC to manage rights.²⁹ As a result, the revenues distributed to artists were substantially increased. Moreover, because of increased competition, JASRAC reduced management costs by as much as 50 percent. In view of those reports, the European Commission has communicated that “the governance and transparency of collective rights management needs to improve and adapt to technological progress.”³⁰

The distortions associated with the administrative inefficiencies and the monopoly power held by collecting societies are enhanced by the oligopolistic structure of the firms operating in the creative industries. As pointed out by Towse (2003), royalty payments to all but the top artists are typically small and firms in the creative industries are typically large, making for a very unequal bargaining situation.³¹ This author argues that the recent policies tending to enhance copyright protection have triggered an increase in merger incentives among the dominant oligopolistic firms in creative industries. According to Towse (2003), the longer copyright lasts, the greater the chance that the market for the work will change in ways that could not be predicted at the time of the original contract. In consequence, extending the copyright length increases oligopolistic firms’ incentives to merge in order to pool the risks associated with the uncertainty of artistic works’ future revenues.³² In contrast, because

²⁹ In particular, management fees for interactive diffusions charged by JASRAC went down from 18 to 11 percent.

³⁰ European Commission Communication, *A Digital Agenda for Europe* (May 19th, 2010). Available at https://europa.eu/european-union/file/1497/download_en?token=KzfSz-CR.

³¹ Towse (2001) provides extensive evidence showing that intermediaries’ share in the revenues far exceeds those of creators.

³² According to the extensive investigation by Bettig (1996) the incentives to accumulate copyright assets seem to be one of the crucial determinants of the intensive process of mergers and acquisitions in the media and entertainment industry that began in the 1970s.

at the beginning of their professional careers individual creators only have their human capital as collateral, they cannot obtain much income if they have to sell their copyrights in order to finance basic living costs. Therefore, as a result of increased copyright term the bargaining power of oligopolist firms tends to increase at the expense of creators, which reduces creative incentives.³³ Moreover, the structure of most Western collecting societies reinforces this effect. In particular, Kretschmer (2003) argues that the joint membership of publishers with creators in collecting societies implies that, because of the stronger bargaining power of publishers, the interests of the creators are underrepresented in these societies. As an illustrative example of the historical conflict of interests between publishers and creators, this author explains the origins of the first German composers' association (GDT) under the chairmanship of Richard Strauss, in 1903. According to Kretschmer (2003), the formation of this composers' interest group was an attempt to counter moves by some publishers to set up an agency for the collective exploitation of musical performance rights. However, in 1916 the publisher Hugo Bock eventually challenged Strauss with the establishment of a rival society: GEMA, which is currently the most important collecting society in Germany.³⁴

3.6 The Role of New Business Models

In view of the theoretical and empirical evidence showing the lack of effectiveness of current copyright policies as a way of promoting creativity, different authors have suggested substantial reforms to the system. As in the case of the regulation of patents, some of those authors have even argued that copyright mechanisms (such as copyright levies and copyright term) are not necessary for providing creative incentives. The idea is that, instead of copyright protection, firms should adopt new business models intended to obtain compensation for the fixed costs associated with the production of creative products. In particular, Varian (2005) has analyzed the way in which different forms of price discrimination can allow creators to obtain the necessary revenues to maintain their creative incentives without the need for antipiracy policies.

In an extensive research, Gopal et al. (2006) have developed this idea further by combining the informative role of sampling, in the digital music industry, with the possibilities of price discrimination. They show that in the presence of online music sampling, uniform pricing for all music items is a suboptimal strategy. Instead, the authors explore a wide range of possibilities for the design of appropriate pricing schemes based on the information on music valuations generated by consumers.

More recently, Nguyen, Dejean and Moreau (2014) have investigated the way in which firms in the music industry can take advantage of the complementarities associated with free

³³ This argument is consistent with the analysis by Rothenbuhler and Dimmick (1982) showing that concentration in the cultural industries leads to a lack of diversity in cultural products.

³⁴ In another historical example, the promotion of the first Spanish author's association (SAE) by the music composer Ruperto Chapí, in 1899, was intended to improve authors' bargaining position by weakening the monopoly power of publisher Florencio Fiscowich. As reported by Young (2006, pp. 272–278), Fiscowich controlled the works of most of Spain's most prominent composers except one, Chapí, who had managed to retain control over his own works. Finally, after a hard battle (including Fiscowich pressures on theatrical impresarios to boycott the play of Chapí's works) Fiscowich was defeated by the authors' association and he agreed to sell his archive to the SAE for the sum of 300,000 pesetas. By early 1902, the SAE could claim victory and fully begin its avowed mission of protecting the rights of authors and composers to their own works. Ironically, the SGAE, which is the current version of the original SAE promoted by Chapí, has become a quasi-monopolist entity mainly dominated by the publishers.

streaming. Using data from a French survey carried out among French Internet users in 2011, these authors show that free music streaming (where the consumer does not possess the music but only has access to it) has no significant effect on CD sales and positively affects live music attendance.

In a similar spirit Bakhshi and Throsby (2014) develop an interesting recent quasi-field experiment, involving the Royal National Theatre's live broadcast of theatre to digital cinemas in the U.K. Those authors report that live broadcasts create greater audiences at the theatre. Therefore, the quasi-field experiment results show that audiences at the theatre are complementary, rather than substitutive, of live broadcasting. The authors conclude that the new business models associated with the development of digital technologies do not necessarily reduce the box-office revenues.

4 GLOBALIZATION AND POLITICAL ECONOMY ISSUES AROUND INTELLECTUAL PROPERTY

In his insightful analysis of the current internationalization process, Stiglitz (2006) has emphasized that intellectual property is one of the most important aspects of globalization, especially as the world moves toward a knowledge economy. In a related paper (Stiglitz, 2008) the same author considers the efficiency and distributive effects of the regulation of intellectual property, paying particular attention to the international aspects of this issue. According to Stiglitz (2008), one of the main gaps that separates developed and developing countries is the disparity in knowledge. This author argues that lowering intellectual property barriers is crucial in helping developing countries to access to the knowledge they need for their economic development and for obtaining cheap pharmaceutical products. However, according to Stiglitz, the main forces driving the configuration of the international regulation of intellectual property are not helping to close that gap. In particular, based on his own experience as a political advisor, Stiglitz explains that most of the people who understand the issues around intellectual property shared his opposition to Trade-Related Aspects of Intellectual Property Rights agreement (TRIPS), which was part of the Uruguay Round of trade negotiations in 1994. In contrast, some of the most important groups interested in this issue (e.g., pharmaceutical, software, and entertainment industries) argued that the stronger the intellectual property rights the better. According to Stiglitz, as a result of the pressure by these interest groups, the TRIPS imposed an unbalanced intellectual property regime.³⁵

The conflicting interests between developed and developing countries regarding intellectual property have been analyzed theoretically in the interesting paper by Helpman (1993). This author assumes a dynamic general equilibrium model in which the North (developed countries) invents new products and the South (developing countries) imitates them. Helpman (1993) shows that if the rate of imitation is small then a tightening of intellectual property harms both the South and the North. Otherwise, there is a conflict of interest between the North and the South: if the rate of imitation is sufficiently large then tightening intellectual property benefits the North and hurts the South. Therefore, the South is always interested in relaxing property rights, but if the rate of imitation is large (imitation is easy) then the North is interested in imposing a tighter intellectual property regime. Interestingly, these theoretical

³⁵ See World Bank (1999).

results seem to be consistent with some political economy explanations of the growing trend to enhance intellectual property in the global economy.

In particular, Shadlen, Schrank and Kurtz (2005) have investigated the empirical evidence of what they call “the new international political economy of intellectual property.” Motivated by the manifest political influence of major software firms in many of the world’s largest economies³⁶ these authors investigate the effects of international obligations on national levels of intellectual property protection in the software industry (which are measured in terms of software “piracy”). To this end, they use data for 80 countries in the period from 1994 to 2002. The authors consider two main international factors: countries’ multilateral obligations under the TRIPS, and bilateral pressures from the U.S.A., in the form of bilateral investment treaties (BITs),³⁷ to increase the protection of intellectual property. Their results provide evidence indicating that membership in the WTO and bilateral pressures from the U.S.A. lead to substantial increases in levels of protection in rich and poor countries. The authors emphasize that the political pressure from the U.S. business constituencies pushed many developing countries to sign bilateral agreements implying higher standards of intellectual property protection than those initially established under the TRIPS.

The global political influence of the large pharmaceutical firms is also remarkable. As in the case of the software industry, the pressure associated with this influence has triggered an increase in patent protection, which is having serious consequences on the welfare of the population in developing countries. As pointed out by Stiglitz (2008), one of the most inefficient (and unfair) consequences of TRIPS was to restrict access to generic medicines, putting these drugs out of the financial reach of most people in developing countries. In particular, the contribution by Collins-Chase (2008) has focused on the particularly harmful effects of the bilateral free trade agreements (FTAs) on those developing countries suffering the HIV/AIDS epidemic. As explained by those authors, the political influence of pharmaceutical firms in the developed countries implies that those FTAs are linked to a stringent enforcement of intellectual property. As a result, many people in those countries have no access to the drugs needed for the treatment of HIV/AIDS.

5 CONCLUSIONS AND FINAL COMMENTS

As indicated in previous sections, it seems that there is robust theoretical and empirical evidence indicating that the growing trend to reinforce patents and copyrights is not justified by its economic consequences. In fact, our revision of the literature suggests that relaxing intellectual property protection would increase innovative and creative incentives as well as welfare.

Moreover, the discussion of the literature dealing with the international aspects of intellectual property reinforces this view, indicating that rather than welfare and international equity considerations, the huge increase in patent and copyright protection seems to be mainly motivated by the political pressure of economically powerful and well-organized patent and copyright holders. The pharmaceutical industry is perhaps one of the most representative

³⁶ See Sell (2003, ch. 5)

³⁷ The crucial role of bilateral agreements with the U.S.A. is justified by the fact that approximately 75 percent of the world’s packaged software is produced in this country. As explained by Shadlen et al. (2005), in those bilateral investment agreements, the U.S.A. includes intellectual property obligations that go beyond TRIPS.

examples of how the current intellectual property system is creating large inefficiencies by different channels, including artificial scarcity (with millions of people having no access to basic medication) and huge wasteful costs in terms of expensive litigation, rent-seeking, lobbying activities,³⁸ and marketing campaigns.³⁹

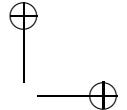
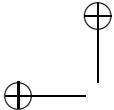
Due to the inefficiencies associated with the current intellectual property system, many reputed authors have suggested alternative reward mechanisms for generating innovative and creative incentives. In particular, Stiglitz (2008) argues that a mixed system more based on incentive schemes such as prizes and grants might be less distortionary and more effective in reducing wasteful costs (in lobbying, litigation and rent-seeking) and in promoting creativity and innovativeness. As previously said, other authors, like Boldrin and Levine (2008, 2012), are even more drastic on this issue and propose the complete elimination of the patent and copyright systems as they currently exist.

REFERENCES

- Adelman, M. (1987), "The New World of Patents Created by the Court of Appeals for the Federal Circuit," *University of Michigan Journal of Law Reform*, 20: 979–1007.
- Adler, M. (1985), "Stardom and Talent," *American Economic Review*, 75(1): 208–212.
- Adler, M. (2006), "Stardom and Talent," in V.A. Ginsburgh and D. Throsby (eds), *Handbook of the Economics of Art and Culture, Volume 1*, Amsterdam: North-Holland, pp. 1: 895–906.
- AEVAL – National Agency for the Evaluation of Public and Quality of Services (2008), *Evaluation of the System of Collective Management of Copyright and Related Rights*, Madrid: Spanish Ministry of the Presidency, available at: <http://www.aeval.es/comun/pdf/evaluaciones/E12eng.pdf>.
- Aghion, P., Bloom, N. and Blundell, R. et al. (2005), "Competition and Innovation: An Inverted U Relationship," *Quarterly Journal of Economics*, 120: 701–728.
- Ahn, I. and Yoon, K. (2009), "On the Impact of Digital Music Distribution," *CESifo Economic Studies*, 55: 306–325.
- Akerloff, G., Arrow, K. and Bresnahan et al. (2002), *Amici Curiae in Support of Petitioners in the Supreme Court of the United States, Eldred versus Ashcroft*, available as Technical Report 01618, Harvard Law School.
- Albinsson, S. (2013), "Swings and Roundabouts: Swedish Music Copyrights 1980–2009," *Journal of Cultural Economics*, 37: 175–184.
- Alcalá, F. and González-Maestre, M. (2010), "Copying, Superstars, and Artistic Creation," *Information Economics and Policy*, 22(4): 365–378.
- Alcalá, F. and González-Maestre, M. (2012), "Artistic Creation and Intellectual Property: A Professional Career Approach," *Journal of Economics and Management Strategy*, 21(3): 633–672.
- Anderson, C. (2006), *The Long Tail: Why the Future of Business is Selling Less of More*, New York: Hyperion Books.
- Arai, Y. and Kinukawa, S. (2014), "Copyright Infringement as User Innovation," *Journal of Cultural Economics*, 38 (2): 131–144.
- Arrow, K. (1962), "Economic Welfare and the Allocation of Resources for Inventions," in R. Nelson (ed.), *The Rate and Direction of Inventive Activity*, Princeton, NJ: Princeton University Press.
- Bae, S.-H. and Choi, J. (2006), "A Model of Piracy," *Information Economics and Policy*, 18: 303–320.
- Bakhshi, H. and Throsby, D. (2014), "Digital Complements or Substitutes? A Quasi-field Experiment from the Royal National Theatre," *Journal of Cultural Economics*, 38: 1–8.
- Banerjee, D. (2003), "Software Piracy: A Strategic Analysis and Policy Instruments," *International Journal of Industrial Organization*, 21: 97–121.
- Banerjee, D. (2006), "Lobbying and Commercial Software Piracy," *European Journal of Political Economy*, 22: 139–155.
- Bar-Isaac, H., Caruana, G. and Cuñat, V. (2012), "Search, Design, and Market Structure," *American Economic Review*, 102(2): 1140–60.

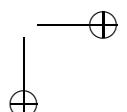
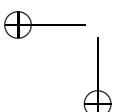
³⁸ The huge political influence of the pharmaceutical industry is reflected in its position as the top lobbyist industry in the U.S.A. with an expenditure of more than 230 millions U.S. dollars. See <http://www.statista.com/statistics/257364/top-lobbying-industries-in-the-us/>.

³⁹ According to Stiglitz (2008), the pharmaceutical industry invests more in marketing than in R&D.



388 *Handbook of game theory and industrial organization: applications*

- Belleflamme, P. and Peitz, M. (2012), "Digital Piracy: Theory," in M. Peitz and J. Waldfogel (eds), *The Oxford Handbook of the Digital Economy*, Oxford: Oxford University Press.
- Berlin, N., Bernard, A. and Fürst, G. (2015), "Time Spent on New Songs: Word-of-mouth and Price Effects on Teenager Consumption," *Journal of Cultural Economics*, 39: 205–218.
- Bessen, J. (2003), "Patent Thickets: Strategic Patenting of Complex Technologies," available at <http://ssrn.com/abstract=327760> or <http://dx.doi.org/10.2139/ssrn.327760>.
- Bessen, J. and Hunt, R. (2007), "An Empirical Look at Software Patents," *Journal of Economics and Management Strategy*, 16(1): 157–189.
- Bessen, J. and Maskin, E. (2009), "Sequential Innovation, Patents, and Imitation," *The RAND Journal of Economics*, 40: 611–635.
- Bessen, J. and Meurer, M. (2009), *Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk*, Princeton, NJ: Princeton University Press.
- Bessen, J. and Meurer, M. (2014), "The Direct Costs from NPE Disputes," *Cornell Law Review*, 99 (2): 387–424.
- Besen, S., Kirby, S. and Salop, S. (1992), "An Economic Analysis of Copyright Collectives," *Virginia Law Review*, 78(1): 383–411.
- Bessen, J., Meurer, M. and Ford, J. (2011). "The Private and Social Costs of Patent Trolls," *Regulation*, 34(4): 26–35.
- Bettig, R. (1996), *Copyrighting Culture*, Boulder, CO: Westview Press.
- Bhattacharjee, S., Gopal, R. and Lertwachara, K. et al. (2007), "The Effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts," *Management Science*, 53 (9), 1359–1374.
- Boldrin, M. and Levine, D. (2004), "Rent Seeking and Innovation," *Journal of Monetary Economics*, 51: 127–160.
- Boldrin, M. and Levine, D. (2008), *Against Intellectual Monopoly*, Cambridge, UK: Cambridge University Press.
- Boldrin, M. and Levine, D. (2012), *The Case Against Patents, FRB of St. Louis Working Paper No. 2012-035A*.
- Boldrin, M., Correa, J., Levine, D. and Ornaghi, C. (2011), "Competition and Innovation," *Cato Papers on Public Policy*, 1: 109–158.
- Burk, D. and Lemley, M. (2009), *The Patent Crisis: And How The Courts Can Solve It*, Chicago, IL: The University of Chicago Press.
- Caves, R. (2000), *Creative Industries: Contracts Between Art and Commerce*, Cambridge, MA: Harvard University Press.
- Chisholm, D. (2004), "Two-part Share Contracts, Risk and Life Cycle of Stars: Some Empirical Results from Motion Picture Contracts," *Journal of Cultural Economics*, 28: 37–56.
- Cohen, L., Gurun, U. and Kominers, S. (2014), "Patent Trolls: Evidence from Targeted Firms," *Working Paper 15-002*, Harvard Business School.
- Collins-Chase, C. (2008), "The Case Against TRIPS-Plus Protection in Developing Countries Facing AIDS Epidemics," *University of Pennsylvania Journal of International Law*, 29(3): 763–802.
- Connolly, M. and Krueger, A. (2006), "Rockonomics: The Economics of Popular Music," in V.A. Ginsburgh and D. Throsby (eds), *Handbook of the Economics of Art and Culture, Volume 1*, Amsterdam: North-Holland, pp. 667–719.
- Crain, W. and Tollison, R. (2002), "Consumer Choice and the Popular Music Industry: A Test of the Superstar Theory," *Empirica*, 29(1): 1–9.
- Ferreira, J. (2010), *Compensation for Private Copying: An Economic Analysis of Alternative Models*, report for Hewlett-Packard, Madrid: ENTER–IE Business School.
- Galasso, A. and Schankerman, M. (2010), "Patent Thickets, Courts, and the Market for Innovation," *The RAND Journal of Economics*, 41(3): 472–503.
- Galasso, A. and Schankerman, M. (2014), "Patents and Cumulative Innovation: Causal Evidence from the Courts," *NBER Working Paper*, June 2014.
- Gayer, A. and Shy, O. (2006), "Publishers, Artists, and Copyright Enforcement," *Information Economics and Policy*, 18: 374–384.
- Gomes, N., Cerqueira, P. and Almeida, L. (2015), "A Survey on Software Piracy Empirical Literature: Stylized Facts and Theory," *Information Economics and Policy*, 32: 29–37.
- Gopal, R., Bhattacharjee, S. and Sanders, G. (2006), "Do Artists Benefit from Online Music Sharing?" *The Journal of Business*, 79(3): 1503–1533.
- Hall, B. and Ziedonis, R. (2001), "The Patent Paradox Revisited: An Empirical Study of Patenting in the US Semiconductor Industry, 1979–95," *RAND Journal of Economics*, 32: 101–128.
- Hall, B. and Ziedonis, R. (2007), "An Empirical Analysis of Patent Litigation in the Semiconductor Industry," *Working Paper*, Department of Economics, University of California, Berkeley.
- Handke, C. (2006), "Plain Destruction or Creative Destruction? Copyright Erosion and the Evolution of the Record Industry," *Review of Economic Research on Copyright Issues*, 3(2): 29–51.
- Handke, C. (2010), *The Creative Destruction of Copyright – Innovation in the Record Industry and Digital Copying*, doctoral dissertation, Erasmus University, Rotterdam.
- Handke, C. (2012), "Digital Copying and the Supply of Sound Recordings," *Information Economics and Policy*, 24: 15–29.



- Heller, M. (1998), "The Tragedy of the Anticommons: Property in the Transition from Marx to Markets," *Harvard Law Review*, 111(3): 621–688.
- Heller, M.A. and Eisenberg, R. (1998), "Can Patents Deter Innovation? The Anticommons in Biomedical Research," *Science*, 280: 698–701.
- Helpman, E. (1993), "Innovation, Imitation, and Intellectual Property Rights," *Econometrica*, 61: 1247–1280.
- Henkel, J. and Von Hippel, E. (2005), "Welfare Implications of User Innovation," *Journal of Technology Transfer*, 30, 73–87.
- Henry, M. and Turner, J. (2006), "The Court of Appeals for the Federal Circuit's Impact on Patent Litigation," *Journal of Legal Studies*, 35: 85–117.
- Huang, K.-W. and Png, I. (2010), "Who Makes the Law? Political Economy Analysis and Evidence from Copyright Levies," manuscript, National University of Singapore.
- Jaffee, A. and Lerner, J. (2004), *Innovation and Its Discontents*, Princeton, NJ: Princeton University Press.
- Jell, F. and Henkel, J. (2010), "Patent Portfolio Races in Concentrated Markets for Technology," *DRUID Working Paper* No. 10-23.
- Katz, A. (2005), "The Potential Demise of Another Natural Monopoly: Rethinking the Collective Administration of Performing Rights," *Journal of Competition Law and Economics*, 1(3): 541–593.
- Kesan, J. and Gallo, A. (2009), "The Political Economy of the Patent System," *North Carolina Law Review*, 87: 101–179.
- Khan, Z. (2004), "Does Copyright Piracy Pay? The Effects of US International Copyright Laws on the Market for Books, 1790–1920," *NBER Working Paper* No. 10271.
- Kim, J.-H. (2007), "Strategic Use of Copyright to Deter Entry," *The B.E. Journal of Economic Analysis and Policy*, 7(1): Article 47.
- Kim, J. (2013), "A Simple Model of Copyright Levies: Implications for Harmonization," *International Tax and Public Finance*, 20(6): 992–1013.
- Kretschmer, M. (2003), "Copyright Societies Do Not Administer Individual Property Rights: The Incoherence of Institutional Traditions in Germany and the UK," in R. Towse (ed.), *Copyright in the Cultural Industries*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 140–161.
- Kretschmer, M. (2009), "Joint Press Release by European Academics (11 March 2009) on the Proposed Directive for a Copyright Term Extension," available at <http://www.cippm.org.uk/downloads/Press%20Release%20Copyright%20Extension.pdf>.
- Kretschmer, M. (2011), *Private Copying and Fair Compensation: An Empirical Study of Copyright Levies in Europe. Project Report*. Newport, UK: Intellectual Property Office.
- Kretschmer, M. and Hardwick, P. (2007), *Authors' Earnings from Copyright and Non-copyright Sources: A Survey of 25,000 British and German Writers*. London/Bournemouth: ALCS/CIPPM.
- Krueger, A. (2005), "The Economics of Real Superstars: The Market for Rock Concerts in the Material World," *Journal of Labor Economics*, 23(1): 1–30.
- Landes, W. and Posner, R. (1989), "An Economic Analysis of Copyright Law," *Journal of Legal Studies*, 18(2), 325–363.
- Landes, W. and Posner, R. (2003a), *The Economic Structure of Intellectual Property Law*, Cambridge, MA: Harvard University Press.
- Landes, W. and Posner, R. (2003b), "Indefinitely Renewable Copyright," *The University of Chicago Law Review*, 70, 471–518.
- Landes, W. and Posner, R. (2004), *The Political Economy of Intellectually Property Law*, Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.
- Lanjouw, J. and Lerner, J. (1996), "Preliminary Injunctive Relief: Theory and Evidence from Patent Litigation," *NBER Working Paper* No. W5689.
- Legros, P. and Ginsburgh, V. (2013), "The Economics of Copyright Levies on Hardware," *Review of Economic Research on Copyright Issues*, 10(1): 20–35.
- Lerner, J. (2009), "The Empirical Impact of Intellectual Property Rights on Innovation: Puzzles and Clues," *The American Economic Review P&P*, 99(2): 343–348.
- Liebowitz, S. and Margolis, S. (2005), "Seventeen Famous Economists Weigh in on Copyright: The Role of Theory, Empirics, and Network Effects," *Harvard Journal of Law & Technology*, 18: 435–457.
- Llanes, G. and Trento, S. (2012), "Patent Policy, Patent Pools, and the Accumulation of Claims in Sequential Innovation," *Economic Theory*, 50(3): 703–725.
- Martínez-Sánchez, F. (2010), "Avoiding Commercial Piracy," *Information Economics and Policy*, 22: 398–408.
- Mehra, S. (2002), "Copyright and comics in Japan: Does Law Explain Why All the Cartoons My Kid Watches are Japanese Imports?" *Rutgers Law Review*, 55: 155–204.
- Merges, R. (1997), *Patent Law and Policy: Cases and Materials*, 2nd edition. Charlottesville, VA: The Michie Company.
- Montoro-Pons, J. and Cuadrado-García, M. (2011), "Live and Prerecorded Popular Music Consumption," *Journal of Cultural Economics*, 35: 19–48.

- Mortimer, J., Nosko, C. and Sorensen, A. (2012), "Supply Responses to Digital Distribution: Recorded Music and Live Performances," *Information Economics and Policy*, 24: 3–14.
- Murray, F. and Stern S. (2007), "Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge? An Empirical Test of the Anti-commons Hypothesis," *Journal of Economic Behavior and Organization*, 63: 648–687.
- Nguyen, G., Dejean, S. and Moreau, F. (2014), "On the Complementarity Between Online and Offline Music Consumption: The Case of Free Streaming," *Journal of Cultural Economics*, 38(4): 315–330.
- Nielsen SoundScan (2008), website, available at <http://www.nielsen.com/us/en/solutions/measurement/music-sales-measurement.html>.
- Novos, I. and Waldman, M. (1984), "The Effects of Increased Copyright Protection: An Analytic Approach," *Journal of Political Economy*, 92: 236–246.
- Oberholzer-Gee, F. and Strumpf, K. (2009), "File Sharing and Copyright," *Innovation Policy and the Economy* 10(1), 19–55.
- Peitz, M. and Waelbroeck, P. (2005), "An Economist's Guide to Digital Music," *CESifoEconomic Studies*, 51: 359–428.
- Peitz, M. and Waelbroeck, P. (2006), "Why the Music Industry May Gain from Free Downloading – The Role of Sampling," *International Journal of Industrial Organization*, 24: 907–913.
- Pitt, I. (2010), "Superstar Effects on Royalty Income in a Performing Rights Organization," *Journal of Cultural Economics*, 34: 219–236.
- Png, I. and Wang, Q. (2009), "Copyright Law and the Supply of Creative Work: Evidence from the Movies," *Working Paper*, National University of Singapore.
- Quinn, G. (2011), "The Cost of Obtaining a Patent in the US," *IPWatchdog.com*, January 28.
- Rosen, S. (1981), "The Economics of Superstars," *American Economic Review*, 71(5), 845–858.
- Rothenbuhler, E. and Dimmick, J. (1982), "Popular Music: Concentration and Diversity in the Industry, 1974–1980," *Journal of Communication*, 32(1), 143–149.
- Salganik, M., Dodds, P. and Watts, D. (2006), "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, 311(5762): 854–856.
- Scherer, F. (2008), "The Emergence of Musical Copyright in Europe 1709 to 1850," *Review of Economic Research on Copyright Issues*, 5(2): 3–18.
- Sell, S. (2003), *Private Power, Public Law: The Globalization of Intellectual Property Rights*, Cambridge, UK: Cambridge University Press.
- Shadlen, K., Schrank, A. and Kurtz, M. (2005), "The Political Economy of Intellectual Property Protection: The Case of Software," *International Studies Quarterly*, 49 (1): 45–71.
- Shapiro, C. (2001), "Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard-Setting," *Innovation Policy and the Economy*, 1, 119–150.
- Stiglitz, J. (2006), *Making Globalization Work*, New York: W.W. Norton & Company, pp. 103–132.
- Stiglitz, J. (2008), "Economic Foundations of Intellectual Property Rights," *Duke Law Journal*, 57: 1693–1724.
- Takeyama, L. (2003), "Piracy, Assymmetric Information and Product Quality," in W.J. Gordon and R. Watt (eds), *The Economics of Copyright. Developments in Research and Analysis*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Towse, R. (2001), *Creativity, Incentive and Reward: An Economic Analysis of Copyright and Culture in the Information Age*, Cheltenham, UK and Northampton, MA, USA.
- Towse, R. (2003), "Copyright and Cultural Policy for the Creative Industries," in O. Granstrand (ed.), *Economics, Law and Intellectual Property*, Boston/Dordrecht/London: Kluwer Academic Publishers, pp. 419–438.
- Tschmuck, P. (2002), "Creativity Without a Copyright: Music Production in Vienna," in R. Towse (ed.), *Copyright in the Cultural Industries*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 210–220.
- Varian, H. (2005), "Copying and Copyright," *Journal of Economic Perspectives*, 19(2): 121–138.
- Waldfogel, J. (2012), "Digital Piracy: Empirics," in M. Peitz and J. Waldfogel (eds), *The Oxford Handbook of the Digital Economy*, Oxford: Oxford University Press.
- Walls, W. (2005), "Modeling Heavy Tails and Skewness in Film Returns," *Applied Financial Economics*, 15: 1181–1188.
- Williams, H. (2013), "Intellectual Property Rights and Innovation: Evidence from the Human Genome," *Journal of Political Economy*, 121: 1–27.
- World Bank (1999), *World Development Report 1998/1999: Knowledge for Development*, available at <https://openknowledge.worldbank.org/handle/10986/5981>.
- Young, C. (2006), Zarzuela: Or Lyric Theatre as Consumer Nationalism in Spain, 1874–1930, PhD thesis, UC San Diego Permalink: <http://eprints.cdlib.org/uc/item/80f3623g>.

APPENDIX: A SIMPLE GENERAL MODEL OF INTELLECTUAL PROPERTY

In this Appendix, we analyze a general stylized model of intellectual property protection to illustrate the idea that the optimal policy on patents and copyright can be formulated in terms of two main types of frictions in the market: (i) those affecting imitation/copying activities and (ii) those affecting the optimal working of patents and copyrights.

The degree of intellectual property protection is defined by the variable $h \in [0, 1]$, with $h = 0$ being the case of zero patent/copyright protection and $h = 1$ being the case of the maximum degree of protection. We assume that a representative innovator/creator can develop a research/creation with social value A and innovation/creation cost given by $K(A, h) = \frac{\gamma}{2}A^2 + \lambda hA$, where $\gamma > 0$ is a technological parameter reflecting the relationship between the social value of innovation and the innovation/creation cost, and $\lambda \in [0, 1]$ is a parameter reflecting the frictions associated with intellectual property. As explained in the discussion of the literature in Section 2, in the case of patents, those frictions might include, among others, those associated with the *patent thickets*, the patent trolls, litigation costs, or portfolio patent races. In the case of copyrights, the detailed discussion in Section 3 suggests that λ might reflect, among others, the informational frictions associated with restricted file-sharing or file-downloading, as well as the effects of different types of entry barriers, such as superstars' increased incentives to invest in promotion and marketing costs, rent-seeking, and lobbying activities. We assume that the innovator/creator's revenue is given by

$$R(A, h) = (\alpha + \mu h)A, \text{ where } \mu, \alpha \in [0, 1]. \quad (\text{A1})$$

In this formulation, μ is a parameter measuring the effectivity of intellectual property to increase innovative/creative incentives, while α is a parameter reflecting the degree of frictions associated with imitation/copying of original goods. As discussed in the main text, those frictions might be due to a wide range of reasons, including short-run capacity limitation, higher production costs, or lower quality of the product provided by the imitators or pirates. Note that, according to this formulation of the model, those frictions imply that innovation incentives appear even in the absence of patents or copyrights. In other words, α is the degree of the innovator's ability to appropriate the social value of the innovation if $h = 0$. According to the previous expression, this degree of appropriation might be enhanced by increasing patent/copyright protection.

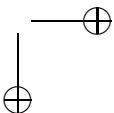
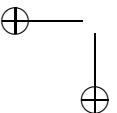
Therefore, the innovator's profit is given by

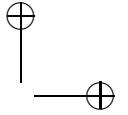
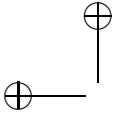
$$\pi(A, h) = R(A, h) - K(A, h) = (\alpha + (\mu - \lambda)h)A - \frac{\gamma}{2}A^2. \quad (\text{A2})$$

We assume that the welfare level in the model is given by

$$W(A, h) = A - K(A, h) = (1 - \lambda h)A - \frac{\gamma}{2}A^2. \quad (\text{A3})$$

Therefore, from the point of view of the static efficiency the welfare loss associated with patents is given by $\lambda h A$. In order to ensure that the rest of the agents in the economy obtain a positive welfare, it must be the case that $W(A, h) - \pi(A, h) = (1 - \alpha - \mu h)A > 0$ for any





$h \in [0, 1]$, which implies $1 - \alpha - \mu > 0 \Leftrightarrow \alpha + \mu < 1$. This is equivalent to assuming that even under the highest degree of intellectual property protection the innovator's ability to appropriate the social value of the innovation is limited. In particular, this might be due to the limited possibilities for implementing price discrimination.

To investigate the optimal intellectual property policy, let us consider a two-stage game where, at the first stage, the welfare-maximizer regulator chooses the degree of patent/copyright protection, h , and then the innovator chooses its level of quality, A , which determines the level of profit and welfare. The innovator's optimal investment in quality, at the second stage, is determined by the first-order condition of profit maximization, which implies that equilibrium level of quality, at that stage, is given by $A^*(h) = \frac{1}{\gamma}(\alpha + (\mu - \lambda)h)$, which substituted in the expression for $W(A, h)$ yields:

$$W^*(h) = \frac{1}{\gamma} \left((\alpha + (\mu - \lambda)h)(1 - \lambda h) - \frac{1}{2}(\alpha + (\mu - \lambda)h)^2 \right). \quad (\text{A4})$$

To investigate the optimal level of intellectual property protection, h^0 , chosen by the regulator at the first stage, let us consider the first derivative of expression (14.4):

$$\frac{\partial W^*(h)}{\partial h} = \frac{1}{\gamma} ((\mu - \lambda) - \alpha\mu - (\mu + \lambda)(\mu - \lambda)h).$$

There might be two cases:

Case 1: If $\mu - \lambda > 0$ then $\frac{\partial^2 W^*(h)}{\partial h^2} = -(\mu + \lambda)(\mu - \lambda) < 0$. Therefore, the welfare function is strictly concave in h and there are three possibilities:

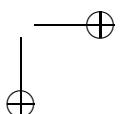
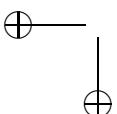
- (Ia) If $\frac{\partial W^*(1)}{\partial h} = \frac{1}{\gamma}(\mu - \lambda)(1 - \alpha\mu - (\mu + \lambda)(\mu - \lambda)) > 0 \Leftrightarrow \alpha < \frac{(\mu - \lambda)(1 - \mu - \lambda)}{\mu}$ then $h^0 = 1$.
- (Ib) If $\frac{\partial W^*(0)}{\partial h} = \frac{1}{\gamma}((\mu - \lambda) - \alpha\mu) > 0$ and $\frac{\partial W^*(1)}{\partial h} < 0 \Leftrightarrow \alpha \in (\frac{(\mu - \lambda)(1 - \mu - \lambda)}{\mu}, \frac{\mu - \lambda}{\mu})$ then h^0 is given by the condition

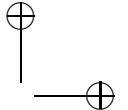
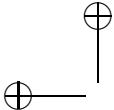
$$\frac{\partial W^*(h)}{\partial h} = 0 \Leftrightarrow h^0 = \frac{\mu - \lambda - \alpha\mu}{(\mu - \lambda)(\mu + \lambda)} \in (0, 1).$$

The comparative statics of this case, with respect to the two frictions, are given by:

$$\begin{aligned} \frac{\partial h^0}{\partial \alpha} &= \frac{-\mu}{(\mu - \lambda)(\mu + \lambda)} < 0, \\ \frac{\partial h^0}{\partial \lambda} &= \frac{-(\mu - \lambda)^2 - 2\lambda\alpha\mu}{((\mu - \lambda)(\mu + \lambda))^2} < 0. \end{aligned}$$

- (Ic) If $\frac{\partial W^*(0)}{\partial h} < 0 \Leftrightarrow \alpha > \frac{\mu - \lambda}{\mu}$ then optimal policy implies $h^0 = 0$.





Case 2: If $\mu - \lambda < 0$ then $\frac{\partial^2 W^*(0)}{\partial h^2} = -(\mu + \lambda)(\mu - \lambda) > 0$. Therefore, the welfare function is strictly convex in h and the optimal h might be given by either $h = 0$ or $h = 1$. Easy calculations yields

$$W^*(0) - W^*(1) = \frac{1}{\gamma} \left(\lambda\alpha - (\mu - \lambda) \left(1 - \frac{1}{2}(\mu + \lambda) \right) \right) > 0,$$

which shows that in this case the optimal degree of intellectual property protection is given by $h^0 = 0$.

The following result summarizes the welfare properties of the model:

Proposition 1 *The optimal level of intellectual property protection, h^0 , is determined as follows:*

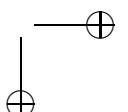
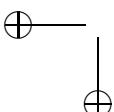
Case 1: If $\mu - \lambda > 0$ there are three subcases:

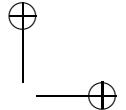
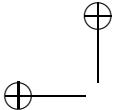
- (1a) *If $\alpha < \frac{(\mu-\lambda)(1-\mu-\lambda)}{\mu}$ then $h^0 = 1$.*
- (1b) *If $\frac{(\mu-\lambda)(1-\mu-\lambda)}{\mu} < \alpha < \frac{\mu-\lambda}{\mu}$ then $h^0 = \frac{\mu-\lambda-\alpha\mu}{(\mu-\lambda)(\mu+\lambda)} \in [0,1]$, which is strictly decreasing in α and λ .*
- (1c) *If $\frac{\mu-\lambda}{\mu} < \alpha$ then $h^0 = 0$.*

Case 2: If $\mu - \lambda < 0$ then $h^0 = 0$.

Therefore, according to Cases 1(c) and 2 in Proposition 1, the greater the degree of frictions, λ and α , the more likely is that the optimal policy involves the complete abolition of intellectual property. In particular, Case 2 corresponds to the case in which the frictions associated with patents/copyrights are so large that the level of innovation, $A^*(h)$, is maximized at $h = 0$. To see this, note that $A^*(h) = \frac{1}{\gamma}(\alpha + (\mu - \lambda)h) < A^*(0) = \frac{1}{\gamma}\alpha$ for any $h > 0$ if $\mu - \lambda < 0$.

Similarly, for the intermediate levels of frictions, associated with Case 1(b), the comparative statics confirms the intuitive idea that the greater the degree of frictions, the lower the level of patent/copyright protection should be.





15. Healthcare and health insurance markets

*Pau Olivella**

1 INTRODUCTION

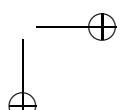
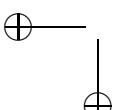
Healthcare services could naively be viewed as any other good like apples or pears. This would mean that any theory that has been developed for private goods could equally be applied to such services. However, healthcare is obviously a very special type of good. It shares with education the fact that it is both a consumption and an investment good. It also shares with education the empirical regularity that both its supply and demand are heavily regulated, at least in most western economies. Nevertheless, there are a few phenomena that are more pervasive in healthcare than in education, let alone other goods: asymmetric information, uncertainty about future demand, and mediated demand. Although uncertainty is also present in many other consumption goods (like transportation, credit, and safety, to mention a few), in health this uncertainty is (more) multidimensional, due to the many different health risks and corresponding health services involved (oncology, traumatology, mental health, and so on). By mediated demand, I mean that in many instances it is not the patient but a physician who makes treatment decisions. Some other features could be especially relevant for healthcare, or at least have been included in the theoretical analysis of healthcare markets. One is the presence of partially altruistic decision makers, in particular doctors, who are trained to act in the clinical interest of the patient. In the same vein, healthcare services are often provided by non-profit institutions that may compete with for-profit firms, a feature that is again in common with education.

For the purposes of this *Handbook*, I will concentrate on the aspects of healthcare demand and supply that consider health as a consumption good. I will also concentrate on those aspects where the tools of industrial organization and game theory are most useful. For this reason, the reader will find it useful to read a few other chapters of this *Handbook* before reading this one, unless the reader has a solid background in the aforementioned tools.

To organize exposition, I make use of the “timing of the healthcare game” depicted in Figure 15.1. I will follow backward induction to order the successive sections. In each section I will discuss some of the solutions that can be found in the literature. To simplify matters, I take several institutional and environmental characteristics as given. First, a set of players is already in place, including providers of health goods and services (hospitals, physicians, pharmaceutical firms); insurers; and individuals with different present and future healthcare needs. I also take the regulatory framework as given.

The first movers are the insurers and, if it exists, a public healthcare provider (like the national health system in the UK or Spain). Insurers decide on the quality of each of the medical services that they commit to cover, subcontract with healthcare providers, and offer insurance contracts. The public healthcare provider also chooses the quality of the services

* The author wishes to thank Inés Macho, David Pérez, Luigi Siciliani and Tom McGuire for their comments to an earlier draft. Any remaining errors are my sole responsibility. The author is affiliated to MOVE and CODE.



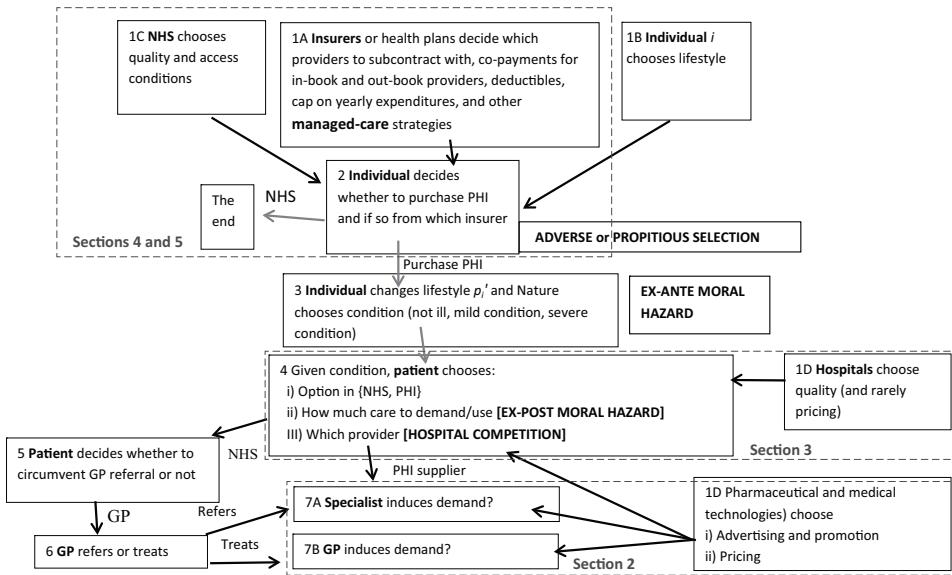


Figure 15.1 The timing of the healthcare game

it covers. Individuals then either choose one of the insurer's contracts or turns to the public provider, if that option is open. They may even go uninsured, as has been the case for a large portion of the population in the USA. In this choice, individuals take into account their expected future health needs, their preference for health, and their risk tolerance. These parameters may be known with better precision by the individual than by the insurer (introducing asymmetric information). Once, and if, the individual is already insured, he may change his life habits, which could lead to a change in his future health needs (ex ante moral hazard). Also, because his spot price of using some healthcare services is decreased due to insurance, his demand for such services may increase (ex post moral hazard). Finally, because some of his consumption decisions are mediated by a physician who could have different objectives from the patient and/or the insurer, these decisions could be distorted, and more so if third parties modify these objectives.

I have marked in discontinuous box-lines the subgames that I will focus on in the successive sections. I have also labeled these boxes with the corresponding section number. (For the pharmaceutical market I will augment the timing with a previous stage where research and development decisions are taken.) The choice of subgames has been driven by my focus on the market mechanism as the main allocation mechanism, rather than the study of bilateral relations and bargaining (although I will mention some works on bilateral negotiations in passing). Some degree of competition between suppliers will be present in every section. This competition will in many instances be limited by regulation. I will provide some final thoughts in the concluding section.

2 THE MARKET FOR PHARMACEUTICAL DRUGS

I begin by describing the extensive-form game that encompasses the interaction between producers (pharmaceutical firms), consumers (patients), payers (insurers, patients), consumers'

agents (doctors), and the planner (the health authority). As with the main healthcare game, some institutional features that are hard to change for political reasons are taken as given (but can be changed when performing policy analysis):

- The doctor prescribes but does not sell drugs.
- Intellectual property is protected for both incumbent and entrant pharmaceutical drugs.
- The government does not fund the research and development of new drugs.
- The regulation path for the approval of a new drug is given.
- The patient does not pay for the drug price in full (copayments or caps on annual expenditures exist).

The literature has been devoted to analyzing the effects of changing one or more of these environmental characteristics.¹ Given these institutional features, the order of moves is generally the following:

1. The pharmaceutical company's (the lab henceforth) takes some research decision: should innovation effort aim at obtaining a me-too drug (incremental innovation) or a blockbuster (radical innovation)? How intense should this effort be?
2. The lab's extra funding decision: is venture capital called-in?
3. The lab's trial decision: how many resources should be devoted to clinical trials?
4. The lab's disclosure/no disclosure of trial results decision.
5. The lab's market access decision: should the firm delay entry into some markets?
6. The lab's pricing decision (in some countries the lab and the payer bargain over price and volume).
7. The health authority's decision whether to list the drug for reimbursement (to subsidize the drug price).
8. The lab's advertising and promotion decision.
9. The game that the patient and the doctor (as a prescriber) play in the choice of treatment.
10. The game that the patient and the pharmacist² play in the choice of treatment.
11. Patient's compliance with the treatment.

Attempting to solve the whole game is of course an enormous enterprise. However, some authors have successfully solved partial views of this game. I will present here in detail the model proposed by González, Macho-Stadler, and Pérez-Castrillo (2016), GMP henceforth, which encompasses stages 1, 6 (prices set by lab), and 9. The trial and disclosure stages (stages 3 and 4) have been studied by Dahm, González, and Porteiro (2015), who show that harsher enforcement of disclosure rules (higher fines for hiding adverse trial results or more intense inspections) are not necessarily beneficial since they may discourage information seeking by the lab. The entry and negotiation stages (stages 5, 6, and 7) have been studied by García-Mariñoso, Jelovac, and Olivella (2011), who show how price regulation that is based on other countries' drug prices (external referencing) affects price negotiations in the latter countries. Advertising and promotion (both direct to consumer advertising and direct to

¹ See Scott-Morton and Kyle (2012) for a survey.

² Or the doctor herself, mainly in Asia and in underdeveloped countries.

physician promotion, or “detailing”, stages 8 and 9) have also been thoroughly analyzed (see Straume, 2014, for a survey).³

Needless to say, the study of the market for pharmaceutical products has greatly benefited from the tools of industrial organization. From a static point of view, drugs are both horizontally and vertically differentiated. From the dynamic point of view, R&D decisions are greatly affected by current market conditions and regulations. Regulation affects the market by limiting entry of new drugs (the approval process can take several years). Prices are often directly regulated as well, or even negotiated between large buyers and the industry, which has prompted the use of the tools of bargaining theory (Bardey, Bommier, and Jullien, 2010).

2.1 A Model of Pharmaceutical Competition and Innovation

As mentioned, GMP solve the game given by stages 1, 6, and 9 of the above timing. The basic features of the model are the following. A continuum of patients exists. Patients are parametrized by the “location” x of their optimal drug in the space of possible drugs, where x is distributed uniformly on the interval $[0, 1]$. A drug is defined by a pair of characteristics: $(\hat{x}, \hat{h}) \in [0, 1] \times R_+$, where the first captures the horizontal position of the drug and the second captures the gross effectiveness of the drug, with $h = 0$ meaning that the drug has the same effect as no treatment. Letting $\ell > 0$ denote the health costs of consuming a less-than-perfect drug (e.g., side-effects), the health gain of a patient of type x_i when drug (x, h) is prescribed is

$$b(h - \ell|x - x_i|),$$

where b is the marginal utility of being healthy. Denoting by p the price of the drug, the benefit to the health system is given by

$$H(x, \hat{x}, \hat{h}, \hat{p}) = b(\hat{h} - \ell|\hat{x} - x|) - \hat{p}.$$

A pharmaceutical firm, say firm 0, is already commercializing drug (x_0, h_0) at a price equal to marginal cost c_0 (for instance, because otherwise a generic drug would enter the market as well). Another pharmaceutical firm, firm 1, has discovered a new drug (x_1, h_1) with associated marginal cost c_1 . Firm 1 freely chooses the price p_1 . Finally, the physician chooses which of the two drugs to prescribe. The authors assume that the doctor is a perfect agent for the planner. That is, the doctor maximizes a utilitarian welfare function that includes consumer surplus and profits. This allows the authors to avoid any agency considerations. It also implies that whether drugs are subsidized or not becomes irrelevant, as the doctor takes into account the full price of the drug, be it paid by the planner or by the patient.

The following notation becomes useful:

$$\begin{aligned}\Delta_x &= |x^1 - x^0|, \\ \Delta_h &= h^1 - h^0, \\ \Delta_c &= c^1 - c^0, \\ \Delta_y &= \Delta_h - \frac{1}{b}\Delta_c.\end{aligned}$$

³ Empirically, pharmaceutical firms engage in much more intense advertising (20 percent of sales approximately) as compared to the general average (between a 4 percent and 5 percent of sales). See Dave (2014).

The last expression is interpreted as the cost-effectiveness of the new drug: it increases with the vertical difference in quality and decreases with the difference in unit costs.

2.1.1 Market shares

The authors show that there exist two thresholds for the price of the entrant:

- p^{\max} , above which the new drug is inviable;
- p^{\min} , below which the old drug is inviable.

Then, for $p \in (p^{\min}, p^{\max})$ some but not all patients are prescribed the new drug. In the case that the incumbent drug is located to the left of the entrant drug, all patients to the right of \tilde{x} are prescribed the new drug, with

$$\tilde{x} = x^0 + \frac{\ell\Delta_x - \Delta_h}{2\ell} + \frac{p^1 - c^0}{2b\ell}.$$

Otherwise, all patients to the left of \tilde{x} are prescribed the new drug, with

$$\tilde{x} = x^0 - \frac{\ell\Delta_x - \Delta_h}{2\ell} - \frac{p^1 - c^0}{2b\ell}.$$

The new drug will tend to (horizontally) compete with the existing drug in a subset of the market. Suppose that the new drug is to the right of the existing drug (i.e., if $x_1 > x_0$). Then the size of this subset is given by $M = 1 - x_0$. For instance, if x_0 is close to one and the new drug is (so unlucky to) fall to the right of x_0 , then the new drug will face extremely fierce competition and M will be close to zero. In that case the new firm will be forced to sell its drug at a relatively low price. If x_0 is close to zero, instead, then the new firm will be able to sell the drug to a large market even if the price is high. A symmetric intuition holds if $x_1 < x_0$.

2.1.2 Pricing decisions

As for the optimal pricing, the most interesting insights (see Propositions 1 and 2 for a full characterization) are as follows. There exists an intermediate range for values of Δ_y and a sufficiently small value of M such that the solution is fully interior, that is, some but not all the patients are prescribed the new drug and the price of the new drug is below p^{\max} .

The effect on consumer surplus (CS) is also interesting. In the same region of parameters, the CS increases when the new drug is launched. (Of course this is also true when the new drug is so superior in the vertical dimension that it overtakes the whole market.) Moreover, the CS increases with the quality of the new drug, whereas it has an inverted-U shape as a function of the horizontal differentiation. As differentiation increases, a better match between patients' specific needs is accomplished, but at the same time the entrant's market power increases as well, and the entrant's price therefore increases. The first effect dominates when differentiation is low and the opposite is true when differentiation is already large.

The conflict of objectives between firms and the planner are clear. The firm prefers to take over the whole market and would therefore prefer to locate its drug (horizontally) as close as possible to the competitor. In contrast, the planner takes into account the better matching between treatments and patients' needs that is accomplished with some differentiation. In

contrast, as for the vertical dimension, the planner's and the firm's objectives are perfectly aligned: the more quality the better. In the absence of transfers, that is, if patients pay the full costs of drugs, they would like to see the maximal horizontal differentiation. In this case, the innovator firm would not have a market backyard where the full CS can be extracted, and would have to compete for a large segment of the market. This in turn would discipline the entrant's prices. In terms of vertical differentiation, consumers do not benefit from such differentiation because the entrant's market power allows it to keep all the benefits of an increase in quality.

2.1.3 Innovation decisions

The authors propose a stylized model of innovation. If innovation is aimed at mimicking the existing drug (incremental innovation), the innovation is successful with probability $q_{in}(I)$, a function that is increasing in the innovation effort or investment I . That innovation can be vertical, in which case it results, if successful, in a drug with $\Delta_x = 0$ and y^1 is given by a random variable with support $[y^0 - \gamma_{ver}, y^0 + \delta_{ver}]$ where both δ_{ver} and γ_{ver} are small. Innovation can also result in a horizontal innovation, leading to a drug where $\Delta_y = 0$ and x^1 is random with support $[x^0 - \gamma_{hor}, x^0 + \delta_{hor}] > 0$, where both δ_{hor} and γ_{hor} are again small.

If innovation is instead aimed at obtaining radical innovation and is successful, which occurs with probability $q_{ra}(I)$, the result could be a drug where both Δ_y and Δ_x are large: the drug could be quite different from the existing one in both the vertical and the horizontal dimensions. However, the random component of such innovation strategy is very different. The support for x^1 is the interval $[0, 1]$, while the support for y^1 is $[y^0 + \kappa, y^0 + \nu]$, with joint distribution given by $f_{ra}(x^1, y^1)$, and κ could even be negative.

The main result is that, once the decision to go for an incremental innovation has been made, the amount of investment that an unregulated firm would choose is the same as the planner's optimum. In contrast, a conflict of interest occurs if the aim is to go for a radical innovation. In this case, the firm underinvests as compared to the planner's optimum. The main intuition is that the firm is able to capture the added financial surplus under an incremental innovation. This is not the case with radical innovations. Hence the stakes of the firm are lower in the latter case.

Does the firm aim for a radical or an incremental innovation? The authors characterize the cases where the firm ceases to aim for a radical innovation whereas the planner would find it beneficial to induce such innovation. More generally, the authors prove that in every case where the planner would prefer an incremental innovation, the firm would never prefer a radical innovation.

The authors extend the analysis to the very plausible case where the doctor does not internalize the full costs of the system but only those that the patient bears. Two reimbursement systems are analyzed. In the first one, the patient pays only a proportion α of the price. This turns out to lead to a very simple extension of the original model. If the marginal costs of the original and the new drug are the same, no allocative effects appear, only the distribution of surplus changes. The price of the new drug increases as α decreases. This is the typical effect of lowering consumers' demand elasticity. If the marginal costs of the original drug are smaller than those of the new drug then an additional, allocative, effect emerges: as copayment α decreases, the new drug is prescribed too often.

In the second reimbursement system analyzed, the patient pays in full the difference between the price of the new and the original drug. Such regulation is called “(internal) reference pricing”, where the price of the incumbent drug acts as the reference. In the case that the new drug involves a different active principle but is aimed at the same type of illness, one speaks of “therapeutic reference pricing”. As compared to the planner’s optimum, in this case the price of the new drug is lower, the new drug is more often prescribed, and the incentives for the firm to bet on radical innovations increase. These results are in line with informal claims of the effects of reference pricing.

3 HOSPITALS: COMPETITION IN QUALITY

Once the individual becomes ill and thus requires some specific health services he will choose between a set of healthcare providers. This choice involves, in our setting, a few specific aspects. First, the individual’s choice does not imply a payment if he is insured or is eligible for healthcare services provided by the government. In the case that he does, this so-called copayment is usually exogenously fixed by regulation. This means that hospitals do not often compete in prices, but in other aspects, like quality. Second, it has been empirically shown that geographical differentiation plays an important role in the choice of hospital (Kessler and McClellan, 2000). Third, the choice is usually mediated by a referring physician (usually a general practitioner) acting as a gate-keeper for specialized healthcare services.⁴ Finally, regulation heavily determines the conditions under which this competition takes place: not only are copayments often regulated, but so are entry, cost reimbursement, and merger activity.

I will address in this section a very specific question. What are the effects on quality of switching from a system where the individual has no choice to a system where the individual does?⁵ The empirical results do not provide a clear answer to this important question.⁶

I will present next the model by Brekke, Siciliani, and Straume (2014), which has many of the special features that I mentioned at the outset, as well as many other realistic assumptions. On the supply side, (i) hospitals hold some degree of market power due to horizontal differentiation; (ii) altruism is an important component in hospital’s preferences; and (iii) increased activity could have learning-by-doing effects leading to decreasing marginal costs of quality as quantity is increased. On the demand side, individuals in good health and with high transportation costs will not seek treatment in any hospital (leading to the existence of local monopolies). Their main contribution is to provide an explanation for the ambiguous empirical evidence for the effect that introducing competition has on observed quality.

⁴ Other examples of limited choice scenarios are closed-book health maintenance organizations (HMOs) or mutualities, which only cover services provided by hospitals listed in the book, or preferred-provider organizations (PPOs), who also have a book of hospitals but allow patients to seek treatment elsewhere at a higher copayment.

⁵ I will not address here other forms of (indirect) competition implemented *via* contests based on relative performance measures (yardstick competition). See Fichera, Nikolova, and Sutton (2014) for a survey.

⁶ Notice that, since insurers contract with hospitals in a given region, many of the insights that we obtained in analyzing competition between health plans carry over to competition between hospitals. However, hospital competition is only relevant once the individual is already ill, which makes the assumption of single service (i.e., single product) competition more justified.

3.1 A Model with a Single Service and Exogenous Prices

A set of n hospitals serve the market. They are equidistantly located on a circle with circumference equal to 1 (Salop's circle). Each hospital receives a price p per patient from a third-party payer. For every hospital $i = 1, 2, \dots, n$, the same quality of treatment q_i is served to all its patients. Patients are uniformly distributed on the circle with total mass equal to 1.

The cost function is the same for all hospitals $C(x_i, q_i)$, where x_i is the number of patients treated at hospital i and q_i measures the overall quality of the hospital. The following assumptions are made on the cost function: $C_x > 0, C_q > 0, C_{xx} \geq 0, C_{qq} > 0$. Very importantly, no assumptions are made on C_{xq} . Hence hospitals' output and quality per patient can be both substitutes through costs ($C_{xq} > 0$) or complements ($C_{xq} < 0$). In the former case, the marginal cost of treating an additional patient increases with quality. Therefore raising quality for all patients would be more expensive the more patients there are to treat. If $C_{xq} = 0$, raising quality for all patients is independent of the number of patients treated, which could indicate that quality is mainly a feature of the hospital at large, like the skills of doctors and nurses. Finally, if the skills of the medical staff are only achieved through training, then having a larger number of patients could lead to smaller costs of raising overall quality, reflecting some degree of "learning by doing".

There are two patient types $\theta \in \{L, H\}$, low and high, differing with respect to two aspects: the gross valuation of treatment and the transportation cost per unit of distance travelled. Namely, for a type θ patient, the valuation is v_θ and the transportation cost is t_θ . Both types are uniformly distributed on the circle with density normalized to 1, with a proportion λ of type H patients. A patient demands either one treatment from the most preferred hospital or no treatment at all. The utility of a patient of type $s \in \{L, H\}$, who is located at x and being treated at hospital i , located at z_i , is given by

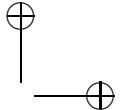
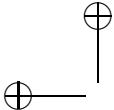
$$U^s(x, z_i) = \begin{cases} V - t_H |x - z_i| + kq_i & \text{if } s = H \\ v - t_L |x - z_i| + kq_i & \text{if } s = L, \end{cases}$$

where q_i is the quality at hospital i and k measures the marginal utility of quality. The authors concentrate on equilibria where the market for H -type patients is fully covered whereas the market for L -type patients is not. Consistently, they refer to the L and H segments as the (local) monopoly and competitive demand segments, respectively.

Hospitals are partly altruistic, with parameter $\alpha \in [0, 1]$ capturing the degree of altruism. The objective function of hospital i is assumed to be given by

$$\pi_i(q_i, \mathbf{q}_{-i}) = \underbrace{T + pX_i(q_i, \mathbf{q}_{-i})}_{\text{Pecuniary}} + \underbrace{\alpha B_i(q_i, \mathbf{q}_{-i})}_{\text{Altruism}} - \underbrace{C(X_i(q_i, \mathbf{q}_{-i}), q_i)}_{\text{Costs}}.$$

Two cases are analyzed: the local monopoly case, where patients willing to attend a hospital are assigned exogenously to the nearest hospital; and the competition case, where adjacent hospitals compete for the high-valuation patients (the H -types).



3.1.1 Case 1: Local monopolies

In this case, hospital i 's demand is

$$X_i(q_i) = \underbrace{\frac{\lambda}{n}}_{\text{Serve the full H segment}} + (1 - \lambda) \cdot 2 \cdot \underbrace{\frac{v + q_i}{t_L}}_{\substack{\text{Demand from L segment} \\ \text{on each side of the hosp.}}}, \quad (15.1)$$

leading to

$$\frac{\partial X}{\partial q_i} = \frac{2(1 - \lambda)}{t_L} > 0 \quad (15.2)$$

and

$$\frac{\partial B_i}{\partial q_i} = X_i(q_i) > 0. \quad (15.3)$$

3.1.2 Case 2: Competition in the H-segment

Assume that both the left and right-adjacent competitors of hospital i set the same quality q_j . Then, hospital i 's demand is

$$X_i(q_i, q_j) = \lambda \cdot 2 \cdot \underbrace{\frac{q_i - q_j + \frac{t_H}{n}}{2t_H}}_{\substack{\text{Demand from H segment} \\ \text{on each side of the hosp.}}} + (1 - \lambda) \cdot 2 \cdot \underbrace{\frac{v + q_i}{t_L}}_{\substack{\text{Demand from L segment} \\ \text{on each side of the hosp.}}}, \quad (15.4)$$

leading to

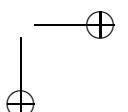
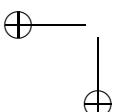
$$\frac{\partial X}{\partial q_i} = \frac{2(1 - \lambda)}{t_L} + \underbrace{\frac{\lambda}{t_H}}_{\text{Additional term}} \quad (15.5)$$

and

$$\frac{\partial B_i}{\partial q_i} = X_i(q_i, q_j) + \underbrace{\frac{\lambda}{t_H} \left(v_H + \frac{q_i + q_j}{2} - \frac{t_H}{2n} \right)}_{\text{Additional term}}. \quad (15.6)$$

3.2 Comparisons

The main result is that there exists a threshold $\hat{\lambda}$ for the proportion of type H patients above which the introduction of competition leads to a smaller equilibrium quality. The intuition is the following. Introducing competition makes the demand of the H-segment elastic to quality. This can be seen by comparing the first term in (15.4) with the first term in (15.1).



Hence raising quality not only increases the altruistic payoff because the benefits enjoyed by inframarginal patients is increased (as it happens under monopoly), but also leads to an increase in the number of patients treated, which in turn raises the hospital's altruistic payoff derived from these marginal patients (compare (15.3) with (15.6)). However, this demand responsiveness also implies that more patients are treated and this can have a negative effect on (pecuniary) profits. This is indeed the case if the marginal mark-up (the one accruing at the marginal patients, $p - C_X$) was negative at the local monopoly solution. This second effect dominates if $p - C_X$ is sufficiently negative, and this occurs if hospitals are sufficiently altruistic. Hence, another result is that higher altruism makes it more likely that competition has negative effects on quality. Indeed, the authors show that the threshold $\hat{\lambda}$ is decreasing in α . In particular, if $\alpha = 0$ so that hospitals are completely selfish, then the mark-up under monopoly is positive and competition will raise equilibrium quality.

Notice that these results hinge on three main assumptions: the existence of decreasing returns to scale in treating patients (favoring negative marginal mark-ups at high levels of activity), sufficient altruism (so that negative marginal mark-ups are consistent with rational decisions on the part of hospitals), and a sufficiently responsive demand function (so that reducing quality does bring cost savings).

The authors also look at two other sources of increased competition: raising the number of hospitals and lowering transportation costs. Also in this case the authors find conditions under which increased competition lowers equilibrium quality.

These results provide possible explanations for the mixed evidence of the effects of increasing competition in the hospital arena.

3.3 Other Explanations

The literature has proposed other possible explanations for this mixed evidence. First, quality may have many different aspects. As in multitasking, some aspects of quality are easier to observe than others, which implies that the introduction of competition will have very different consequences on each of these dimensions. For instance, waiting time data are usually publicized whereas mortality rates in specific services are not. One would therefore expect that the introduction of competition induces a reduction in waiting times but an increase in mortality rates. This insights have been partially confirmed by empirical analyses (Cookson and Dawson, 2006). In the same vein, hotel-related amenities (enjoying a single room, for instance) are easily appraised (Goldman and Romley, 2008), whereas medical quality requires a good understanding of concepts like morbidity or quality-of-life-adjusted survival. Although one could argue that the referring physician can be a good source of advice in these dimensions, geographical proximity and comfort considerations often get in the way of the final choice. Incidentally, these other elements, which could have different weights for different individuals, are often introduced in competition models by means of a random shock to the utility that individuals derive from attending any given hospital, or from enrolling any given health plan – as we will see in the next section.

Some other interesting empirical regularities have been documented. If service prices (that is, the per-service transfer from the third party payer to the hospital) are negotiated instead of set exogenously, increased competition leads to lower instead of higher mortality rates. In the USA, for instance, prices are negotiated between HMOs and hospitals whereas in Medicare, prices are determined by average costs in previous periods (Gowrisankaran and Town, 2003).

Incidentally, a theoretical model that includes price negotiation is bound to be more complex (see Barros and Martinez-Giralt, 2006), and perhaps this is the reason why these observations still lack a theoretical explanation.

4 STATUTORY HEALTH INSURANCE

In statutory health insurance (SHI henceforth, also referred to as “managed competition”, Enthoven, 1978) systems, individuals are allowed to choose their insurer but their contribution to the financing of healthcare is independent of their true health risk. Individuals’ financial contributions go to a common fund that is then used to pay insurers prospectively, that is, insurers receive an amount, usually referred to as *capitation rate*, for each individual who enrolls in the health insurance plan offered by that insurer. These capitation rates are intended to avoid insurers’ incentive to select patients. For instance, if treating a 62-year-old male costs 200 euros per year on average, whereas treating a 32-year-old male costs 100, then the fund pays a capitation rate of around 200 euros for each 61-year-old male enrollee and 100 euros for each 32-year-old male. If the costs are estimated using the expenditures in previous years conditional on observables then one speaks of *conventional risk adjustment* (Glazer and McGuire, 2000). In principle, this system has several advantages. First, since the insurer is the residual claimant, any reductions in treatment costs accrue to the insurer, thus providing the right incentive to contain costs. Second, if capitation rates are correctly risk-adjusted, no risk selection incentives remain. Third, since insurers compete to attract patients, they do not have incentives to skimp on quality.⁷ Designing the payment system to achieve these goals, however, is not a straightforward matter, as we will see below.

A first step in assessing a risk-adjustment system is to determine what would happen in its absence. This could mean two things: either a purely private and unregulated health insurance system is in place, or a capitation system where capitation rates are not risk-adjusted is in place. Before discussing these two alternative regulatory frameworks, one needs to first distinguish between two scenarios. In the first one, insurers and individuals have the same information about the individual’s health risks, or symmetric information scenario. In the second, individuals have privileged information about their health risks, or asymmetric information scenario. I address the former scenario in the next subsection. I devote the next section in full to the latter.

4.1 Private Health Insurance Under Symmetric Information

In this scenario, the individual must pay the insurance premium in full out of pocket, and insurers can condition these premia on the individual’s health risk. In the absence of ex ante

⁷ A more sophisticated system consists in letting insurers ask for some out-of-pocket (OoP) premium from their enrollees. The total revenue per insuree is then the sum of the OoP premium plus the capitation rate. This adds some degree of flexibility to the system. The regulator can then allow different degrees of OoP-premium discrimination. For instance, gender discrimination may be banned while some coarse age categorization may be allowed. The interaction between the information on risks conveyed by the allowed variables and the information conveyed by the variables used for risk adjustment leads to an interesting statistical problem. See McGuire et al. (2013) and Van de Ven et al. (2017). Such a system is often referred to as *risk-equalization*. However, the term “equalization” is also used to distinguish adjustment of capitation rates or premia from the adjustment of fee-for-service payments to average service costs (as in the so-called “dialogistic-related group” payment mechanisms).

and ex post moral hazard, this leads to full insurance (efficient risk sharing) and premia that will depend on expected healthcare costs. Any intervention in this market must have redistributional purposes, since allocative efficiency is already assured. Such interventions are usually aimed at having low risks cross-subsidize high risks, by mandating some degree of “community rating”, which limit the variables that insurers can use to price discriminate. For instance, the EU has recommended that premia be independent of gender (Commission of the European Communities, 2003). Such policies pose two dangers. The first one is that, if insurers make losses when attracting some applicants, they may engage in direct risk selection, that is, denying insurance to these applicants. Therefore, community rating is usually coupled with “open enrollment” rulings, whereby any applicant wishing to sign an offered contract must be accepted. The other danger is that individuals who are in good health status may not even apply, making the whole system financially inviable. This problem can be addressed by imposing “mandatory enrollment”.⁸

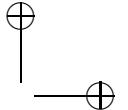
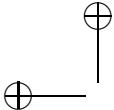
The combination of all of these regulations can still lead to another problem, often referred to as “indirect selection”, where insurers select risks through actions that are difficult to observe by the regulator. For instance, insurers can selectively advertise their health plans to certain groups of individuals (Van de Ven et al., 2017), or may craft the design of supplementary insurance (that is, insurance covering services like dentistry or aesthetic treatments that are not usually included in the main package; Lamiraud, 2014) in order to attract certain risk types. Another form of indirect selection is the so called “service-level selection”, which has been the subject of extensive research in the last two decades. In very basic terms, service-level selection consists in insurers underproviding those services that are usually demanded by those in worse health status. I will present the main ideas behind this selection strategy in the next subsection.

Once mandatory and open enrollment together with community rating are imposed, the situation is similar to a capitation system where individuals do not pay any out-of-pocket premium directly to insurers and where capitation rates are not adjusted to risk. The only difference is the route that money follows. In a private insurance system subject to these regulations, the individual pays a community-rated premium directly to the insurer. In a non-adjusted capitation scenario without out-of-pocket premia, the individual contributes to a common fund and then this common fund pays the capitation rate to the insurer. In the absence of transaction costs and rent-seeking behavior on the part of the managers of the common fund, the two scenarios are equivalent.

4.2 Service-level Selection

Let me start presenting the minimal ingredients with which understand the main trade-offs. There are S health services in the production of health $s \in \{1, \dots, S\} \equiv \mathcal{S}$. For instance, $s = 1$ stands for mental health services, $s = 2$ stands for oncology treatments, $s = 3$ for traumatology services, and so on. In a general model (which I will simplify later on), individuals have a prior probability of contracting a disease or suffering an accident that makes these services necessary. A state of the world ω can be described as the list of all services needed, which is a subset of $\{1, 2, \dots, S\}$. Therefore the set of all possible states of the world Ω is the power set

⁸ For instance, both regulations are present in the state-level market places created by the Affordable Care Act (aka “Obamacare”) that came into play in 2014 in the USA. See McFadden, Noton, and Olivella (2015).



$2^{\mathcal{S}}$. One can then define a probability distribution over $2^{\mathcal{S}}$ to describe the “health status” of an individual, which I denote by x , which I also refer to as an individual’s “type”. Let the set of types be given by $\{1, 2, \dots, X\}$. I also assume that there are n_x individuals of each type x in the population.

I describe the quality or intensity of a care in service s in terms of the amount of money spent on that service for individual of type x , which is denoted by $m_{xs} > 0$. The health benefits that an individual of type x obtains from using a subset of services \mathcal{S}^* is given by $\tilde{v}_x(\{m_{xs}\}_{s \in \mathcal{S}^*})$.

The first simplification that is often exploited in the literature is to assume that, for any $s, s' \in \mathcal{S}$, all individuals will eventually require each service s .⁹ A second simplification is to assume that, for any $s, s' \in \mathcal{S}$ and for any type x , the health benefits that an individual of type x derives from an expenditure m_{xs} in service s are independent of the expenditures in another service s' . This implies that the benefits in each service s and for any type x can be described by a single-valued function $v_{xs}: R_+ \rightarrow R_+$, where it is assumed that v'_{xs} is decreasing.

These two simplifications, taken together, allow us to derive the optimal allocation of resources. It is simply achieved by choosing the expenditure $m \geq 0$ that maximizes $v_{xs}(m) - m$ at each service s and type x . This leads to the allocative efficiency condition $v'_{xs}(m) = 1$ for all x and s . Let $m = m_{xs}^*$ be the expenditure satisfying this condition. If the planner could observe the functions v_{xs} for any type x and service s , she could mandate expenditure m_{xs}^* for each service s and type x in order to implement allocative efficiency. However, these expenditures are not usually directly regulated, possibly because the planner is not able to observe all the valuation functions.

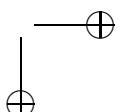
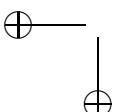
In the absence of this direct regulation, how do insurers and healthcare providers behave? Suppose that the insurer sets a fixed budget \bar{M}_s per patient of type x in each service s . Suppose also that the healthcare providers (e.g., the clinicians subcontracted by the insurer) in each service s act as perfect agents of the patients they are responsible for. Clinicians in service s then choose the expenditure m_{xs} for each type x to maximize

$$\sum_{x=1}^X n_x v_{xs}(m_{xs}) \text{ subject to } \sum_{x=1}^X n_x m_{xs} = \sum_{x=1}^X n_x \bar{M}_s.$$

Let λ_s be the Lagrange multiplier associated with the budget constraint. For any type x , the first-order condition for m_{xs} is $v'_{xs}(m_{xs}) = \lambda_s$. In words, clinicians in any given service s spend resources for each patient so that the marginal valuation is the same across all types. Notice that this condition is a necessary but not sufficient condition for allocative efficiency, since marginal valuations may vary from service to service. As we will see, this condition can also be interpreted as the planner requiring that, at any service s , all individuals face the same access conditions.

It turns out that one can rewrite the insurer’s problem of allocating a budget to each service by means of the so-called “shadow price approach”, where the shadow price is denoted by q . In general, when an individual of type x enjoys expenditures of m_{xs}^0 euros in service s , this entails a marginal valuation equal to $v'_{xs}(m_{xs}^0)$. Then let $q_{xs}^0 = v'_{xs}(m_{xs}^0)$ which we refer to

⁹ Alternatively, one could assume that, for any service $s, s' \in \mathcal{S}$, the events “needing s ” and “needing s' ” are independent. This would allow us to describe the health status of individual type x as a vector $(p_{x1}, \dots, p_{xS}) \in [0, 1]^S$, where p_{xs} is the probability that an individual of type x needs service s .



as the shadow price of one dollar of expenditure in service s . It is then possible to express the allocation of expenditures *as if* an individual's demand for expenditure in service s were always met but the individual faced a price q_{xs}^0 per dollar of expenditure. I will also denote the demand (of expenditures in) service s by type x as $D_{xs}(q)$, i.e., $D_{xs}(q)$ is the solution for m in $v'_{xs}(m) = q$.

Here are some examples. Assume that $s = 1$ stands for mental health care services and $s = 2$ stands for traumatology services. In the absence of any regulation, the insurer could make access to mental health services more difficult for a 65-year-old male (type x) than for a 45-year-old male (type y), that is, the insurer could let $q_{x1} > q_{y1}$. It could also make access to one euro of expenditure in mental health services more difficult than access to one euro of expenditure in traumatology for the same individual (i.e., $q_{x1} > q_{x2}$). Allocative efficiency is attained for an individual of type x at service s if $q_{xs} = 1$. If instead $q_{xs} > 1$ (respectively, < 1) then service s is underprovided (respectively, overprovided) for an individual of type x . Finally, the budget-allocation problem analyzed above leads to $q_{xs} = q_{ys}$ for all x, y and s , but is silent about whether, for any given type x , q_{xs} is equal to $q_{xs'}$ or not in any two services s and s' . Hence, this allocation mechanism could also be the one resulting from an anti-discrimination law. For instance, if $x = \text{women}$ and $y = \text{men}$ then men and women should have the same access conditions to any service s , or $q_{xs} = q_{ys}$ for all s .

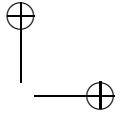
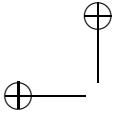
In sum, the shadow price q_{xs} conveys and summarizes how difficult it is to access any given service s by individual of type x .¹⁰ A more general interpretation of the shadow price, and the one that most of the literature has adopted, is that the shadow price summarizes all the “managed care” strategies that are set in order to contain demand. These shadow prices are the basis of competition among insurers.

Let me advance the main results here. Distortions in access (i.e., departures from the first-best allocation where $q_{xs} = 1$ for all x and s) will result from insurers trying to select risks. These distortions, as we will see below, are governed by two forces. One is “predictability”, which measures how able patients are to foresee their future needs. The other is “predictiveness”, which conveys how expenditures in a given service s correlate with an individual's overall expenditures. If predictability is low at service s , insurers will not find it useful to distort the quality of this service, since individuals will not be very responsive to such distortion. If predictability in service s is high, distorting the quality at this service will be an effective tool with which to select risks. In this case, services where predictiveness is positive and strong (negative and strong), that is, where intense usage predicts high (low) overall costs across all services, will be distorted downwards (upwards). All these ideas are formalized next.

4.2.1 Insurers' behavior

Suppose that the health system is financed through general taxation (as in Medicare) or through contributions that are independent of one's type. Under equal access regulation, it

¹⁰ There is one instance where this price becomes an actual price (in the usual sense), namely, when the individual pays a fraction q_{xs} of the expenditures on service s , that is, when q_{xs} is a proportional copayment. A more subtle example of such price is the cost of waiting, although this example would require several assumptions on how waiting time and waiting time costs are related to expenditures. Both examples would bring some additional revenues or cost savings that would have to be taken into account when formulating the insurer's profits.



suffices to describe the vector $q = \{q_1, q_2, \dots, q_S\}$ of shadow prices that the insurer chooses. The first exercise is to determine what the services are that would be distorted due to selection incentives *in the absence of any risk adjustment*. Hence, let r be the compensation per individual, independent of the individual's type, that the insurer gets from the common fund. I follow Ellis and McGuire (2007), EM henceforth, closely in the remainder.¹¹ EM distinguish between the actual needs of service s of an individual of type x , or $v_{xs}(\cdot)$, and the insurer's beliefs about the individual's beliefs about such needs, or $\tilde{v}_{xs}(\cdot)$. I will refer to these needs as "foreseen needs". For instance, $\tilde{v}_{xs}(\cdot)$ could vary across x in a lesser degree than the true $v_{xs}(\cdot)$. This could be the case if, for instance, individuals are not very good at anticipating their future health needs. This will play a very important role in the insurer's incentives to distort the quality of that given service. In accordance, an individual expects to demand $\tilde{D}_{xs}(q_s) \equiv \{\tilde{v}'_{xs}\}^{-1}(q_s)$ when facing shadow price q_s .

Let us assume that there are two insurers, 1 and 2, each choosing a vector of shadow prices, (q_1^1, \dots, q_S^1) and (q_1^2, \dots, q_S^2) . The utility that an individual of type x expects to enjoy by enrolling in plan i has two components, a direct expected utility component given by $\sum_{s=1}^S \tilde{v}_{xs}(\tilde{D}_{xs}(q_s^i))$ and a noise component $-\eta_x^i$. For instance, $-\eta$ could reflect an individual's random transportation costs. We assume that all these noise components are independent and identically distributed (i.i.d.) uniformly in $[0, 1]$. Hence all insurers are ex ante symmetric.

Letting $\mathbf{q}^i = (q_1^i, \dots, q_S^i)$ and $u_x^i(\mathbf{q}^i) = \sum_{s=1}^S \tilde{v}_{xs}(\tilde{D}_{xs}(q_s^i))$, an individual of type x will choose insurer 1 if $u_x^1(\mathbf{q}^1) - \eta_x^1 > u_x^2(\mathbf{q}^2) - \eta_x^2$. Letting $\Delta_x^u(\mathbf{q}^1, \mathbf{q}^2) = u_x^1(\mathbf{q}^1) - u_x^2(\mathbf{q}^2)$, this occurs if $\eta_x^1 - \eta_x^2 < \Delta_x^u(\mathbf{q}^1, \mathbf{q}^2)$. Now, since η_x^1 and η_x^2 are uniformly distributed in $[0, 1]$, the difference $\eta_x^1 - \eta_x^2$ is distributed according to a triangular distribution Φ with modal point 0. Therefore, the probability that this individual chooses insurer 1 is given by $\Phi(\Delta_x^u)$. Notice that in the symmetric equilibrium, where $\Delta_x^u = 0$ for all x , this probability becomes $\Phi(0) = 1/2$, and that $\Phi'(0) = 1$. This will also become important when analyzing plan i 's behavior.

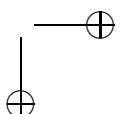
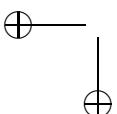
Plan i 's profits are

$$B(q) = \sum_x n_x \Phi\left(\left(\Delta_x^u(\mathbf{q}^1, \mathbf{q}^2)\right)\right) \left[r - \sum_{s=1}^S n_x D_{xs}(q_s^1) \right].$$

Now, in order to compute the incentives to distort quality in service s with respect to allocative efficiency, we take the partial derivative of this profit function with respect to q_s^1 , and then we evaluate it at the symmetric equilibrium and at the first-best shadow price ($q_s = 1$).¹²

¹¹ The main departures from their analysis are: (1) I keep track of the population of individuals of each type; (2) in order to preserve symmetry among plans, I assume that noise affects all insurers equally; and (3) from the outset, I keep track of the difference between the actual insurer's expenditures in each service and the expenditures that the insurer believes that the patient foresees he or she is going to enjoy, whereas EM only make this distinction at the first-order conditions.

¹² Other approaches assume perfect competition and therefore make use of the zero-profit condition to derive predictions on the sign of selection. The advantage of the approach that I present here is that it uses the profit-maximization conditions to derive implications for insurer behavior. These implications can subsequently be used to derive the conditions for optimal risk adjustment.



This partial derivative is given by

$$\begin{aligned} \frac{\partial B}{\partial q_s} &= \sum_x n_x \Phi' \left(\left(\Delta_x^u (\mathbf{q}^1, \mathbf{q}^2) \right) \right) \tilde{v}'_{xs} \left(\tilde{D}_{xs} \left(q_s^1 \right) \right) \tilde{D}'_{xs} \left(q_s^1 \right) \cdot \\ &\quad \left[r - \sum_{s=1}^S n_x D_{xs} (q_s^1) \right] + \sum_x n_x \Phi \left(\left(\Delta_x^u (\mathbf{q}^1, \mathbf{q}^2) \right) \right) \left[-D'_{xs} (q_s^1) \right]. \end{aligned} \quad (15.7)$$

Let us first consider the actual expenses. A common assumption in the literature is that, for any fixed service s , all types have the same elasticity of actual demand with respect to the shadow price q_s and, moreover, this elasticity coincides with the elasticity of foreseen demand. That is, it is assumed that $\eta_{sx} = \tilde{\eta}_{sx} = \eta_s$ for all x and s .¹³ The elasticity of demand for service s is given by $\eta_s = D'_{xs} (q_s^1) \frac{q_s}{D_{xs} (q_s^1)}$. Hence $D'_{xs} (q_s^1) = \frac{\eta_s D_{xs} (q_s)}{q_s}$ can be substituted into (15.7). Also $v'_{xs} (D_{xs} (q_s^1)) = q_s$ (by definition) can be substituted into (15.7). For each individual of type x , let $M_x = \sum_{s=1}^S D_{xs} (q_s^1)$, total expenditures per type x , and let $B_x = r - M_x$, the profit per individual of type x . Let us now consider the foreseen expenditures. We have $\tilde{\eta}_{xs} = \tilde{D}'_{xs} (q_s) \frac{q_s}{\tilde{D}_{xs} (q_s)} = \tilde{D}'_{xs} (q_s) \frac{\tilde{v}'_{xs} (\tilde{D}_{xs} (q_s))}{\tilde{D}_{xs} (q_s)}$ and therefore $\tilde{v}'_{xs} (\tilde{D}_{xs} (q_s)) \tilde{D}'_{xs} (q_s) = \tilde{\eta}_{xs} \tilde{D}_{xs} (q_s) = \eta_x \tilde{D}_{xs} (q_s)$. This allows us to rewrite (15.7) as

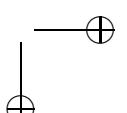
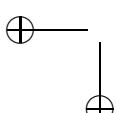
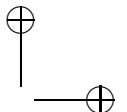
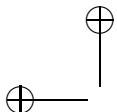
$$\begin{aligned} \frac{\partial B}{\partial q_s} &= \sum_x n_x \Phi' \left(\left(\Delta_x^u (\mathbf{q}^1, \mathbf{q}^2) \right) \right) \eta_s \tilde{D}_{xs} (q_s) B_x \\ &\quad + \sum_x n_x \Phi \left(\left(\Delta_x^u (\mathbf{q}^1, \mathbf{q}^2) \right) \right) \left[-\frac{\eta_s D_{xs} (q_s)}{q_s} \right]. \end{aligned}$$

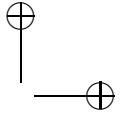
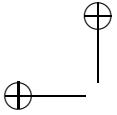
Notice that in this expression I (and EM) still distinguish individuals' forecasts $\tilde{D}_{xs} (q_s)$ from true levels of expenditures $D_{xs} (q_s)$.

Let us now impose symmetry, that is, let $\Delta_x^u (\mathbf{q}^1, \mathbf{q}^2) = 0$ to get $\Phi' (0) = 1$ and $\Phi (0) = \frac{1}{2}$, hence

$$\frac{\partial B}{\partial q_s} = \sum_x n_x \eta_s \tilde{D}_{xs} (q_s) B_x - \sum_x n_x \frac{1}{2} \frac{\eta_s D_{xs} (q_s)}{q_s}.$$

¹³ This is true, for instance if valuation functions $\tilde{v}_{xs} (\cdot)$ and $v_{xs} (\cdot)$ are power functions and types are parametrized by a term that enters multiplicatively. Indeed, suppose that $v_{xs} (z) = \pi_{xs} \mu_s z^{1/\mu_s}$ and $\tilde{v}_{xs} (z) = \tilde{\pi}_{xs} \mu_s z^{1/\mu_s}$. Then actual demand of service s by an individual of type x is obtained by solving $v'_{xs} (z) = \pi_{xs} z^{1/\mu_s - 1} = q_s$, hence $D_{xs} (q_s) = q_s^{\frac{\mu_s}{1-\mu_s}} \pi_{xs}^{\frac{\mu_s}{\mu_s-1}}$, while elasticity of actual demand is given by $\eta_s = -\frac{\mu_s}{1-\mu_s}$, independent of type x . Hence the elasticity of actual demand across individuals for a given service is the same while actual demand may differ across services. As for foreseen expenditures, we have that foreseen demand of service s by an individual of type x is obtained by solving $\tilde{v}'_{xs} (z) = \tilde{\pi}_{xs} z^{1/\mu_s - 1} = q_s$, hence $\tilde{D}_{xs} (q_s) = q_s^{\frac{\mu_s}{1-\mu_s}} \tilde{\pi}_{xs}^{\frac{\mu_s}{\mu_s-1}}$, while the elasticity of foreseen demand is given by $\tilde{\eta}_s = -\frac{\mu_s}{1-\mu_s}$, which is the same as the one obtained for actual expenditures.





410 Handbook of game theory and industrial organization: applications

The next step is to evaluate this expression at $q_s = 1$, which entails $D_{xs}(1) = m_{xs}^*$ and therefore

$$\frac{\partial B}{\partial q_s}(1) = \eta_s \sum_x n_x \tilde{D}_{xs}(1) B_x - \eta_s \sum_x n_x \frac{1}{2} m_{xs}^*.$$

Finally, denoting the total actual expenditures in service s by $\bar{m}_s = \sum_x \frac{1}{2} n_x m_{xs}^*$, we have

$$\frac{\partial B}{\partial q_s}(1) = \eta_s \sum_x n_x \tilde{D}_{xs}(1) B_x - \eta_s \bar{m}_s,$$

which can be rewritten as

$$\frac{\partial B}{\partial q_s}(1) = \bar{m}_s \left(\eta_s \sum_x \frac{n_x \tilde{D}_{xs}(1)}{\bar{m}_s} B_x - \eta_s \right).$$

If this expression is negative (positive) at some s , it means that the insurer has incentives to lower q_s below 1 (increase q_s above 1), i.e., to overprovide (underprovide) that service. EM use the previous expression to define an index I_s that quantifies these incentives:

$$I_s \equiv \frac{\partial B}{\partial q_s} \cdot \frac{1}{\bar{m}_s},$$

which they refer to as the *selection index for service s* . We can write

$$I_s \equiv \frac{\partial B}{\partial q_s} \cdot \frac{1}{\bar{m}_s} = \frac{\eta_s}{\bar{m}_s} \sum_x n_x \tilde{D}_{xs}(1) B_x - \eta_s.$$

Denoting by σ_z the standard deviation of a variable z , and by \bar{z} as its average, we write the correlation between \hat{m}_{xs} and B_x across all types $x = 1, \dots, N$ evaluated at $q_s = 1$ as:

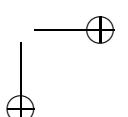
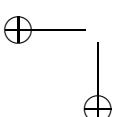
$$\rho_{\hat{m}_s, B} = \frac{\sum_{x=1}^N n_x \tilde{D}_{xs}(1) B_x - \bar{m}_s \bar{B}}{\sigma_{\hat{m}_s} \sigma_B}.$$

Isolating the first term in the numerator and substituting into the index yields

$$I_s \equiv \frac{\eta_s}{\bar{m}_s} (\sigma_{\hat{m}_s} \sigma_B \rho_{\hat{m}_s, B} + \bar{m}_s \bar{B}) - \eta_s.$$

or

$$I_s \equiv \sigma_B \eta_s \left(\frac{1}{\bar{m}_s} \sigma_{\hat{m}_s} \rho_{\hat{m}_s, B} + \frac{\bar{B}}{\sigma_B} - \frac{1}{\sigma_B} \right).$$



Letting $C = -\left(\frac{B}{\sigma_B} - \frac{1}{\sigma_B}\right)$, a term that does not depend on service s , the index is

$$I_s \equiv \sigma_B \eta_s \left(\frac{1}{\bar{m}_s} \sigma_{\hat{m}_x} \rho_{\hat{m}_s, B} - C \right).$$

Finally, since $\pi_x = r - M_x$, we have that $\rho_{\tilde{D}_s, B} = -\rho_{\tilde{D}_s, M}$. Therefore,

$$I_s \equiv \underbrace{-\eta_s}_{\text{Positive}} \underbrace{\sigma_B}_{\text{Predictability}} \begin{pmatrix} \underbrace{\frac{\sigma_{\tilde{D}_x}}{\bar{m}_s}}_{\text{Predictiveness}} & \underbrace{\rho_{\tilde{D}_s, M}}_{-} & C \end{pmatrix}.$$

The index formally combines the two forces driving the incentives to distort service s that I introduced above. Predictability is given by $\frac{\sigma_{\tilde{D}_x}}{\bar{m}_s}$, which is the coefficient of variation of predicted expenditures at service s . This will be rather large if individuals expect to have very different expenditures depending on their type x . Predictiveness is given by $\rho_{\tilde{D}_s, M}$, which conveys to what degree expenditures in service s explain overall expenditures (i.e., total expenditures across all services). If expenditures in a given service s are positively (negatively) correlated to overall expenditures then predictiveness will be positive (negative). If predictability is strong and predictiveness is positive in service s , this will lead to a very positive index of selection for that service. This means that the plan has strong incentives to raise q_s over 1, that is, to underprovide that service. Intuitively, if the plan expects individuals to respond a lot to such distortions because they are able to predict them, and if at the same time these individuals are prone to demanding healthcare services intensely, incentives to avoid such individuals will be strong. The opposite case occurs if, for a service s , predictability is still strong but predictiveness is negative. Finally, if individuals forecast similar expenditures across all types (predictability is weak) then insurers know that any distortions in service quality will be useless in attracting certain types of individuals and pushing away others.

Several authors (EM themselves and McGuire et al., 2014, for instance) have implemented this index to estimate distortion incentives using US data. They consistently find that mental health services are the most prone to downward distortions in quality, as both predictiveness and predictability are strong for these services. The question is then how should premia be risk-adjusted in order to counteract these incentives. Of course if the set of observables available (gender, age, past expenditures) were perfect predictors of future costs, the problem would be closed. Just let $r_x = M_x$. However, this is not the case (observables account for less than 40 percent of the cost variation even in countries using the most sophisticated methods). Hence the problem is to find the weights that each observable should have in the (second-best) risk-adjustment formula (see, for instance, Glazer and McGuire, 2002).

5 ASYMMETRIC INFORMATION

As it has long been recognized, asymmetric information is a pervasive phenomenon in healthcare provision and insurance. I will concentrate here on the market responses to the

presence of asymmetric information, rather than on the issues arising in bilateral relations like doctor/patient, payer/provider, or regulator/insurer.

It is important to distinguish here between two scenarios that may seem equivalent but lead to different issues and call for different tools of analysis. One is where information is truly asymmetric, that is, individuals hold privileged information about their health risks and the industry responds by offering menus of contracts (as in Rothschild and Stiglitz, 1976). The other is where information is symmetric but regulation forbids one of the sides, usually the supply side, to act on part of this information. For instance, under community rating, premia cannot be made to depend on observable individual characteristics that are informative about the individual's future health costs. Similarly, in the previous section I have analyzed the issue of service-level selection in the market of health plans when capitation rates are not adjusted to risk. Although one could see this as a form of asymmetric information, in fact insurers try to select applicants directly (for instance, by outright rejection, usually forbidden) or indirectly (through service-level selection or targeted advertising). Importantly, in most of the literature (including that on service-level selection), insurers are assumed to compete by offering a single health plan for each allowed class of observables and only the symmetric equilibrium outcome is characterized, where all insurers offer the same health plan to each class. It is therefore assumed that neither the individual insurers nor the industry as a whole are trying to screen individuals by means of menus of health plans. This implies that no incentive compatibility constraints need to be imposed.

There are, however, a few works that do address the truly asymmetric information scenario. Lorenz (2015) does allow for market-wide menus of health plans but still assumes that insurers compete in single-contract offers. This means that different insurers aim at attracting different risk types, which constitutes an asymmetric equilibrium. Glazer and McGuire (2000), Jack (2006), and Olivella and Vera-Hernández (2007) instead allow insurers to offer menus of health plans but do concentrate on the symmetric equilibrium where all insurers offer the same menu. This latter distinction deserves a bit more discussion.

5.1 Menu vs Single Contract Competition

Whether menus are implemented at the market scale (in the sense that insurers are restricted to offer a single health plan) or at the insurer level (insurers are allowed to offer menus) turns out to be irrelevant under two conditions: (i) markets are perfectly competitive; (ii) a large enough proportion of individuals with high health risks exists. If one relaxes condition (ii) then there exists a lower bound for the proportion of high risks below which no competitive Nash equilibrium exists (Rothschild and Stiglitz, 1976). Moreover, this lower bound is higher if firms are allowed to offer menus. The literature has proposed several ways to restore existence, and perhaps the most popular one takes the form of the (non-Nash) Miyazaki-Wilson-Spence (MWS) equilibrium notion.¹⁴ The main elements in this equilibrium concept are, first, that firms are indeed allowed to offer menus of contracts. Second, if the proportion of high risks is low then this menu entails low risks cross-subsidizing high risks, that is, insurers make profits from the low risks and losses from the bad risks, but break even on average. This implies that insurers have an incentive to drop the loss-making contract. If one insurer deviates in this fashion then the rest of the insurers make losses, since they will have to absorb the high

¹⁴ See, for instance, Netzer and Scheuer (2014) for a formal definition of this solution concept.

risks displaced by the insurer who deviated. This leads to the third element in the MSW solution concept, namely, it is assumed that, because these other firms are making losses, they too deviate by withdrawing their menus from the market. That is, the MWS allows players to deviate upon another player's deviation. Once these second-round deviations take place, the original deviation becomes unprofitable since again all high risks will now accept the deviant firm's outstanding contract.¹⁵ A nice feature of the MWS equilibrium notion is that the equilibrium menu of contracts always exists and is always second-best efficient. This implies that any intervention by the government may alter the distribution of resources but cannot bring about a Pareto improvement.

Menus also become important if the market is not perfectly competitive, that is, if one relaxes condition (i) above. Under horizontal differentiation, cross-subsidization becomes more intense the lower the proportion of high risks in the population. However, this conclusion does not always require the application of the MWS equilibrium concept. The reason is that firms may not have an incentive to withdraw the contract that attracts the high-risk individuals, for two reasons. First, if differentiation is sufficiently important, then all elements in the equilibrium menu of contracts bring positive profits (even the contract attracting high risks). Second, suppose that differentiation is not strong enough and therefore the contract attracting high risks, say contract α , brings (moderate) losses, and let β be the contract aimed at attracting the low risks. Suppose also that an insurer D (for "deviator") drops contract α . Then, due to the existence of transportation costs, a few of the high risks who would have accepted α will remain with insurer D and accept her contract β . This turns out to bring very large losses for this insurer. This effect may be strong enough to restore the existence of a competitive Nash equilibrium even with moderate transportation costs (Olivella and Vera-Hernández, 2007).

5.2 The Health Insurance Sector

All these are general lessons that apply to any insurance market as long as some conditions are met. The question is then whether they carry over to health insurance. In the next subsections I will discuss some of the extensions of the canonical insurance model that have (or should) be made in order to addressing the specificities of the health sector.

5.2.1 Indemnity vs healthcare provision

First of all, the model used to derive the previous results is one of indemnity insurance, meaning that the insuree receives an ex post pecuniary transfer in case of illness. In health insurance, the individual instead receives treatment. However, one could say that the insurer pays for the healthcare services that the individual would have had to pay in case of falling ill. The problem with this argument is that it presumes that (i) healthcare costs per individual are constant, that is, independent of the number of patients treated, and (ii) all losses in health are insurable.

¹⁵ Wilson's argument was originally formulated for single-contract competition (Wilson, 1977). Some authors have enlarged the game that insurers and individuals play in order to provide a Nash underpinning to this solution concept. See, for instance, Netzer and Scheuer (2014) and Diasakos and Koufopoulos (2015).

5.2.2 Multidimensional screening

Second, the results are derived for a scenario where the individual faces a single source of risk, whereas in reality an individual may require a wide array of different health services, leading to a multidimensional screening problem. This is an extremely tough problem to solve. The main difficulty is that individuals' types cannot be ranked in terms of the degree of their risk. For instance, an individual may have a low risk of suffering lung cancer but a high risk of suffering mental health problems, say he is type LH, while another has the opposite risk characteristics, say he is type HL. Even assuming that the two events (suffering lung cancer and suffering from mental health problems) are independent, one has to deal with four risk types: LL, LH, HL, HH.¹⁶ This implies that it is *a priori* impossible to know which of the potential incentive compatibility constraints (ICCs) are binding. For instance, the set of binding ICCs could either be $\{HH \rightarrow HL, HL \rightarrow LH, LH \rightarrow LL\}$ or $\{HH \rightarrow LH, LH \rightarrow HL, HL \rightarrow LL\}$. The very scarce literature on multidimensional screening in insurance in a competitive setting has used two tricks to overcome this problem. One is to assume that only the two intermediate types exist, that is, *HL* and *LH* (Fluet and Pannequin, 1997); the other is to assume that only two types of individuals exist with different sources of risk (or "perils") and assuming that individuals can self-select by choosing different deductibles for each peril (Crocker and Snow, 2011).

Propitious selection Multidimensionality of types may arise due to other sources of heterogeneity that do not involve health risks per se. The sources of heterogeneity that have received most attention (more empirically than theoretically) include the following: wealth (Wambach, 2000), risk aversion (Smart, 2000; Villeneuve, 2003; Finkelstein and McGarry, 2006; and Cutler, Finkelstein, and McGarry, 2008), cognitive ability (Fang, Keane and Silverman, 2008), and precautionary behavior (De Meza and Webb, 2001). The common idea in all these works is that the same individuals who are prone to purchasing insurance may also be less prone to making use of the insured services. For instance, individuals who are more risk averse tend to have safer life habits. However, it is interesting to note that, under perfect competition among insurers, theory has established that propitious selection cannot occur (Chiappori et al., 2006).

5.2.3 Duplicate vs opt-out private health insurance

Another special characteristic of health insurance is the coexistence, in many countries, of a public health provider that is financed through general taxation and a set of private health insurers. The main question here is whether the individuals who purchase PHI are the high or the low risks. This is important for two reasons. First, vertical equity is compromised if the two sectors offer very different coverage (cross-subsidization is more difficult). Second, financial viability of the sector attracting the high risks may be compromised. The predictions of the theory and the empirical conclusions depend on the following distinction: when an individual decides to purchase PHI, does he or she continue financing (and being eligible

¹⁶ If the two sources of risk are not independent we then have four states of the world in terms of how many services are needed: none, only cancer, only mental health, both. This requires establishing three distinct probabilities, and therefore $2^3 = 8$ types exist even in the simplest case of dichotomous types: LLL, LLH, LHL, ..., HHH. No theoretical model exists dealing with this situation at the time of writing this survey.

for treatment in) the public option? If the answer is yes one speaks of a “substitutive” or “duplicate” PHI insurance system, which can be found in many countries that have a national health system, like Spain, UK, or Portugal. If the answer is no, one speaks of “PHI with opt-out”, as one can find in Germany or Chile. One important stumbling block in comparing these two systems is that many other features of the regulatory environment differ between duplicate and opting-out countries. For instance, in Germany it is not only true that individuals may opt out of the public alternative, but is also true that only individuals with a minimum income per capita are allowed to do so. Moreover, the public insurer in Germany is in fact a set of public insurers rather than a monolithic national health system like the NHS in the UK. In very basic terms, the main conclusion in the asymmetric information scenario is that high risks tend to purchase PHI in duplicate (this conclusion is quite general; see Olivella and Vera-Hernández, 2013) whereas one can construct equilibria where the low risks are the ones opting out of a system where such an option is available (Panthöfer, 2015).

It turns out that most of the intuitions behind these results can be shown by means of figures. Before doing this, I need to establish some basic notation. Let the initial after-tax wealth be the same for all individuals and given by $w > 0$. Upon falling ill, the individual faces a financial loss $L > 0$, also the same for all individuals. This occurs with probability $\pi \in [0, 1]$. An insurance contract is given by a pair of numbers, referred to as premium and coverage: (P, c) . The individual pays P ex ante and receives payment c if he falls ill. An individual’s final wealth if healthy (sick) is $n(a)$. The utility function over money is u , so expected utility is given by $U = \pi u(a) + (1 - \pi) u(n)$. An insurer’s profits are given by $B = w - n(1 - \pi) - \pi a - \pi L$. The no-insurance point (or autarchy point) is $A = (n, a) = (w, w - L)$. Individuals are of two types, H and L, with associated probabilities of falling ill equal to π_H and π_L , with $\pi_H > \pi_L$. It can be shown that, at any contract (n, a) , the indifference curve of the low risk is steeper than for the high risk. Isoprofits are straight lines with slopes $\frac{1-\pi_H}{\pi_H} < \frac{1-\pi_L}{\pi_L}$. These are also the slopes of the indifference curves for each type at full insurance ($n = a$).

The equilibrium under asymmetric information and a duplicate PHI system is depicted in Figure 15.2, where I illustrate the finding that only high risks purchase PHI. The equilibrium pair of contracts that would arise in the absence of a public option is $(\alpha_H^*, \hat{\alpha}_L)$, where the high risk receives a full insurance contract (the same as under symmetric information), whereas the

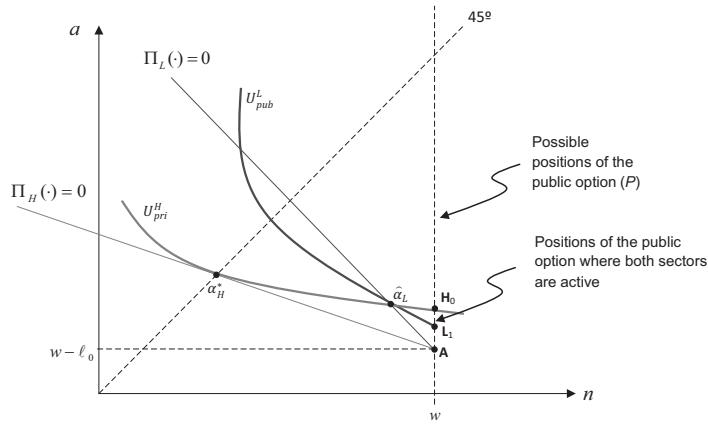


Figure 15.2 The equilibrium under duplicate PHI

low risk receives a contract such that profits are zero and the high risk is indifferent between the two contracts. With a public option and no opt-out, the vertical discontinuous line crossing the autarchy point A is the set of possible positions of the public option. The reason is that w is the after-tax wealth and all individuals contribute independently of whether private health insurance has been purchased or not. Within these possible positions, only the ones between points H_0 and L_1 are compatible with the two sectors being active. For such positions, it is indeed the case that only the high risks purchase PHI, as the low risks prefer the public option to $\hat{\alpha}_L$. However, no comparisons can be made here with the *laissez faire* since w is net of the taxes needed to finance the public sector (see online Appendix D in Olivella and Vera-Hernández, 2013).

A possible equilibrium under a PHI with opt-out under a minimum-wealth requirement is depicted in Figure 15.3. The only difference from the previous model is that now w stands for the initial wealth gross of the contribution that the individual pays if he does not opt out. This allows us to keep track of the initial disposable wealth of the individuals who do opt out (as they are no longer required to contribute). In this equilibrium, the opposite result obtains. Suppose that the unconditional risk mix leads to a pooling zero profit condition given by the line Q. However, eligibility restrictions for opting out do not allow some low risks to opt out, whereas all high risks (irrespective of their eligibility conditions) stay in the public sector. Then the risk mix that the public sector faces is given by the less-steep line Q' . Suppose that the public option is given by point P in line Q' . The remaining low risks do opt out and the competitive forces in the private sector lead to contract α_L^{**} . The high risks are in a better position than in the *laissez faire* due to the forced cross-subsidization existing in the public sector (compare contracts P and α_H^*). This softens the incentive compatibility constraint so all low risks who are allowed to opt out are also better off than in the *laissez faire* (compare α_L^{**} to $\hat{\alpha}_L$). The rest of the low risks are worse off: they are forced to subsidize the high risks (compare P to $\hat{\alpha}_L$). Panthöfer (2015), in a slightly more complex model, shows that this and other equilibria are possible in this setting. Therefore, the direction of selection can only be ascertained by econometric methods. This author uses data from Germany that establishes

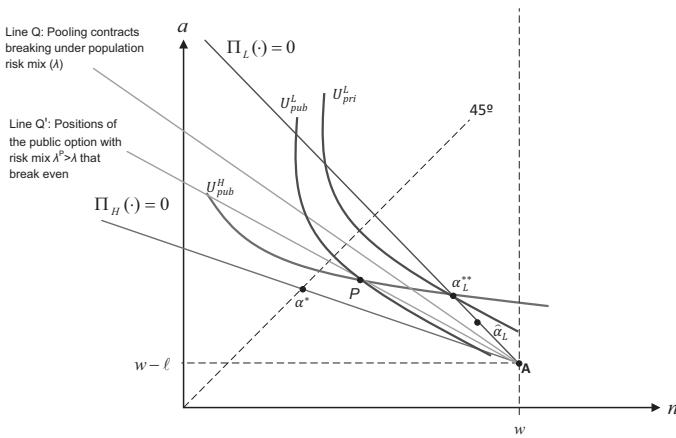


Figure 15.3 A possible equilibrium under PHI with opt-out

that selection into the private insurance sector is adverse, which is consistent with the situation depicted in Figure 15.3.

6 CONCLUSIONS

Health economics is one area in economics where sound theoretical analysis has taken some time to sink in. Before the mid-1980s it was seldom the case that one could find any theoretical modeling in most publications. However, once game and industrial organization theorists (as well as theorists in other fields like macroeconomics, development economics and labor economics) realized the wide array of health economics problems that could benefit from theory, the situation changed fast. Despite the fact that a single model cannot encompass all the elements present in many situations, theory has been useful in guiding empirical analysis and in the production of relevant insights. It has also forced health economists to be more precise in defining the different elements at play (regulation, market structure, preferences, beliefs, and so on).

I have tried to provide a glimpse of theoretical modelling here, although I have concentrated on the market as the main mechanism to allocate resources. The reader who is interested in learning more on the application of the tools in the title of this *Handbook*, and many others, should take a careful look at the textbooks by Barros, and Martinez-Giralt (2012) and Zweifel, Breyer, and Kifmann (2009). These authors do not shy away from the formulation of complex equations and results. The fact that complex theories are making their way into health economics textbooks is a token of the important change in the field I just referred to.

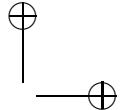
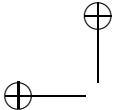
Finally, as can be inferred from reading this chapter, many challenging topics for further research remain.

REFERENCES

- Bardey, D., Bommier, A., and Jullien, B. (2010), "Retail Price Regulation and Innovation: Reference Pricing in the Pharmaceutical Industry", *Journal of Health Economics* 29, 303–316.
- Barros, P., and Martinez-Giralt, X. (2006), "Models of Negotiation and Bargaining in Health Care", in A.M. Jones (ed.), *The Elgar Companion to Health Economics*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Barros, P., and Martinez-Giralt, X. (2012), *Health Economics: An Industrial Organization Perspective*, Boston, MA and New York: Routledge.
- Brekke, K.R., Siciliani, L., and Straume, O.-R. (2014), "Hospital Competition and Quality with Regulated Prices", *Scandinavian Journal of Economics* 113, 444–469.
- Chiappori, P.-A., Jullien, B., Salanié, B., and Salanié, F. (2006), "Asymmetric Information in Insurance: General Testable Implications", *The RAND Journal of Economics* 37, 783–798.
- Commission of the European Communities (2003), "Proposal for a Council Directive Implementing the Principle of Equal Treatment between Women and Men in the Access to and the Supply of Goods and Services", *Council Directive 2003/657*, available at <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52003PC0657>.
- Cookson, R. and Dawson, D. (2006), "Hospital Competition and Patient Choice in Publicly Funded Health Care", in A.M. Jones (ed.), *The Elgar Companion to Health Economics*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Crocker, K., and Snow, A. (2011), "Multidimensional Screening in Insurance Markets with Adverse Selection", *Journal of Risk and Insurance* 78, 287–307.
- Cutler, D., Finkelstein, A., and McGarry, K. (2008), "Preference Heterogeneity and Insurance Markets: Explaining a Puzzle of Insurance", *American Economic Review* 98, 157–162.

- Dahm, M., González, P., and Porteiro, N. (2015), "Monitoring, Punishment and Selective Reporting", mimeo, Universidad Pablo de Olavide, August 22.
- Dave, D.M. (2014), "Pharmaceutical Marketing and Promotion", in T. Culyer (ed.), *Encyclopedia of Health Economics, Vol. 3*, Amsterdam: Elsevier, 9–19.
- De Meza, D., and Webb, D. (2001), "Advantageous Selection in Insurance Markets", *The RAND Journal of Economics* 32, 249–262.
- Diasakos, T., and Koufopoulos, K. (2015), "(Neutrally) Optimal Mechanism Under Adverse Selection: The Canonical Insurance Problem", available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2680295.
- Ellis, R.P., and McGuire, T.G. (2007), "Predictability and Predictiveness in Health Care Spending", *Journal of Health Economics* 26, 25–48.
- Enthoven, A. (1978), "A Consumer Choice Health Plan: A National Health Insurance Proposal Based on Regulated Competition in the Private Sector", *New England Journal of Medicine* 298, 650–658 and 709–720.
- Fang, H., Keane, M.P., and Silverman, D. (2008), "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market", *Journal of Political Economy* 116, 303–350.
- Fichera, E., Nikolova, S., and Sutton, M. (2014), "Comparative Performance Evaluation: Quality", in T. Culyer (ed.), *The Encyclopedia of Health Economics Vol. 1*, Amsterdam: Elsevier, 111–116.
- Finkelstein, A., and McGarry, K. (2006), "Multiple Dimensions of Private Information: Evidence from the Long-term Care Insurance Market", *American Economic Review* 96, 938–958.
- Fluet, C., and Pannequin, F. (1997), "Complete vs. Incomplete Insurance Contracts under Adverse Selection with Multiple Risks", *The Geneva Papers on Risk and Insurance Theory* 22, 81–101.
- García-Mariñoso, B., Jelovac, I., and Olivella, P. (2011), "External Referencing and Pharmaceutical Price Negotiation", *Health Economics* 20, 737–756.
- Glazer, J., and McGuire, T.G. (2000), "Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care", *The American Economic Review* 90, 1055–1071.
- Glazer, J., and McGuire, T.G. (2002), "Setting Health Plan Premiums to Ensure Efficient Quality in Health Care: Minimum Variance Optimal Risk Adjustment", *Journal of Public Economics* 84, 153–173.
- Goldman, D., and Romley, J. (2008), "Hospitals as Hotels: The Role of Patient Amenities in Hospital Demand", *National Bureau of Economic Research Working Paper* 14619.
- González, P., Macho-Stadler, I., and Pérez-Castrillo, D. (2016), "Private versus Social Incentives for Pharmaceutical Innovation", *Journal of Health Economics*, 50, 286–297.
- Gowrisankaran, G., and Town R.J. (2003), "Competition, Payers, and Hospital Quality", *Health Services Research* 38, 1403–1422.
- Jack, W. (2006), "Optimal Risk Adjustment with Adverse Selection and Spatial Competition", *Journal of Health Economics* 25, 908–926.
- Kessler, D., and McLellan, M. (2000), "Is Hospital Competition Socially Wasteful?", *Quarterly Journal of Economics* 115, 577–615.
- Lamiraud, K. (2014), "Switching Costs in Competitive Health Insurance Markets", in T. Culyer (ed.), *Encyclopedia of Health Economics, Vol. 3*, Amsterdam: Elsevier.
- Lorenz, N. (2015), "The Interaction of Direct and Indirect Risk Selection", *Journal of Health Economics* 42, 81–89.
- McFadden, D., Noton, C., and Olivella, P. (2015), "Minimum Coverage Regulation in Insurance Markets", *SERIES* 6, 247–278.
- McGuire, T.G., Glazer, J., and Newhouse, J.P., et al. (2013), "Integrating Risk Adjustment and Enrollee Premiums in Health Plan Payment", *Journal of Health Economics* 32, 1263–1277.
- McGuire, T.G., Newhouse, J.P., and Normand, S.-P. et al. (2014), "Assessing Incentives for Service-level Selection in Private Health Insurance Exchanges", *Journal of Health Economics* 35, 47–63.
- Netzer, N. and Scheuer, F. (2014), "A Game Theoretic Foundation of Competitive Equilibria with Adverse Selection", *International Economic Review* 55, 399–422.
- Olivella, P., and Vera-Hernández, M. (2007), "Competition Among Differentiated Health Plans under Adverse Selection", *Journal of Health Economics* 26, 233–250.
- Olivella, P., and Vera-Hernández, M. (2013), "Testing for Asymmetric Information in Private Health Insurance", *Economic Journal*, 123, 96–130.
- Panthöfer, S. (2015), "Risk Selection Under Public Health Insurance with Opt-Out", *UC3M Working Papers*.
- Rothschild, M., and Stiglitz, J. (1976), "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information", *Quarterly Journal of Economics* 90, 629–649.
- Scott-Morton, F. and Kyle, M. (2012), "Markets for Pharmaceutical Products", in M. Pauly, T. McGuire, and P.P. Barros (eds), *The Handbook of Health Economics, Vol. 2*, Amsterdam: North-Holland.
- Smart, M. (2000), "Competitive Insurance Markets with Two Unobservables", *International Economic Review* 41, 153–169.
- Straume, O.-R. (2014), "Advertising Health Care: Causes and Consequences", in T. Culyer (ed.), *The Encyclopedia of Health Economics, Vol. 1*, Amsterdam: Elsevier, 51–55.

- Van de Ven, W.P.M.M., Van Vliet, R.C.J.A., and Van Kleef, R.C.J.A. (2017), “How Can the Regulator Show Evidence of (No) Risk Selection in Health Insurance Markets? Conceptual Framework and Empirical Evidence”, *European Journal of Health Economics*, 18, 167–180.
- Villeneuve, B. (2003), “Concurrence et Antisélection Multidimensionnelle en Assurance”, *Annales d’Économie et de Statistique* 0(69): 119–142.
- Wambach, A. (2000), “Introducing Heterogeneity in the Rothschild–Stiglitz Model”, *Journal of Risk and Insurance* 67, 579–591.
- Wilson, C. (1977), “A Model of Insurance Markets with Incomplete Information”, *Journal of Economic Theory* 16, 167–207.
- Zweifel, P., Breyer, F., and Kifmann, M. (2009), *Health Economics*, 2nd edition, Berlin/Heidelberg: Springer.



16. The microeconomics of corruption

*Roberto Burguet, Juan-José Ganuza and José G. Montalvo**

1 INTRODUCTION

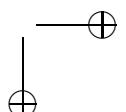
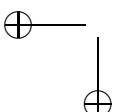
It is widely acknowledged that corruption is one of the most important factors affecting the creation and the distribution of wealth around the world. Nowadays, there is a consensus that corruption not only leads to redistribution of surplus, but also generates many distortions in the economy. Its consequences range from poverty to lack of investment or poor education indicators. Corruption ranks high on the list of concerns not only for the public opinion but also for governments and international organizations. Thus, national policies, but also supranational efforts to curb corruption abound. A leading example is the OECD's "Convention on Combating Bribery of Foreign Public Officials in International Business Transactions", but similar projects have been launched by the World Trade Organization, the United Nations, and the Council of Europe.

Economists have not ignored this concern. The last decades have witnessed a surge of research in economics, both theoretical and empirical, with corruption as its main issue. This chapter surveys in a structured fashion some of the main lines of that research. On the theory side, our emphasis will be on how microeconomic analysis, including game theory and mechanism design, have contributed to our understanding of the mechanisms of corruption and to the design of public policies devoted to fight it. On the empirical side, we will review empirical and experimental literature that helps our understanding of the economic impact of corruption, as well as test the validity of the theoretical results and the assumptions under which they are derived.

Our starting point is the standard model of corruption based on a principal–supervisor–agent (client) setting, which we discuss in Section 2. In this model, the principal – society, the public, or its representatives – delegates some task to a supervisor. The supervisor's decisions affect third parties or clients – citizens, entrepreneurs, regulated firms. Corruption arises when these third parties and the supervisor – official – collude to take the “wrong” decision; that is, a decision that is not optimal for the principal, but one the client prefers. We discuss optimal delegation contracts that take into account this possibility of collusion (corruption). Preventing corruption is costly, and so one of the questions is, under what conditions is it optimal to prevent corruption completely?

The core of corruption is the agreement between the official and the third party. In Section 3, we focus on the details that characterize the negotiations leading to these agreements and

* Roberto Burguet gratefully acknowledges the financial support of the Spanish Ministerio de Economía y Competitividad under project ECO2014-59959-P and the Generalitat de Catalunya (Grant 2014 SGR 510). Juan-José Ganuza gratefully acknowledges the financial support of the Spanish Ministerio de Economía y Competitividad under project ECO2014-59225-P, the Barcelona GSE Research Network, and the Generalitat de Catalunya. José G. Montalvo acknowledges the financial support from the project ECO2014-55555-P from the Spanish Ministerio de Economía y Competitividad, the Barcelona GSE Research Network and the Generalitat de Catalunya (ICREA-Academia Fellowship and Grant 2014 SGR 546).



their stability. Apart from the issues that affect any sort of negotiations – e.g., asymmetric information – corrupt agreements face an additional complication: their illegal nature. That is, parties cannot count on the instruments of law – and its courts – for enforcement. The deal faces the “hold-up” problem, and this in turn may be used by the principal to prevent or hinder corruption.

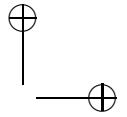
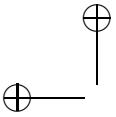
After discussing the three-tier, basic model of delegation and corruption, we open the lens to include in the field of vision the context in which corrupt deals take place. Thus, in Section 4 we review the trade-off between the costs of corruption and the market failures that justify regulation or, in general, public intervention. The focus is then switched to bureaucracy, its size, its compensation, incentives and selection.

An even wider angle is needed to discuss the deeper sources of corruption. In this regard, it is often argued that the existence of rents that can be illegally appropriated is a prerequisite for the existence of corruption. Also, lack of competition in the market generates rents. Therefore, improving competition is a recipe for reducing rents and, it is argued, this also leads to less corruption. Rose-Ackerman (1996) summarizes this idea when she writes: “In general any reform that increases the competitiveness of the economy helps reduce corrupt incentives.” We review this allegation in Section 5. We also discuss competition on the supply side, that is, among bureaucrats. Indeed, the literature has discussed several ideas on how to organize the bureaucracy – one-stop, sequential approval, job rotation, etc. – in order to minimize the impact of corruption. We will survey some of these discussions and those related to how market structure and bureaucracy structure may interact.

Section 6 will close the loop in theoretical discussion by returning to contract selection and exploring public procurement in depth. A recent literature has discussed at some length how corruption affects the performance of given procurement mechanisms. Here the relationship between corruption and competition is also central. Indeed, bribery may not only be an instrument for rent extraction, but also a device that facilitates collusion. Nevertheless, the causality may also go in the opposite direction: the number of competitors affects the incidence of corruption. Moreover, these relationships are often subtle. Also in Section 6, we will discuss a few papers that study the design of an optimal procurement mechanism in the presence of potentially corrupt agents.

Many theoretical aspects will be left outside the scope of this chapter, some of them central to the legal and sociological analysis of corruption – culture, social norms, trust – some also widely discussed in development economics and political economy. However, before turning to the empirical literature, we will briefly discuss one of them: corruption as a problem of – society’s – multiple equilibria. Corruption and the rule of law may be two alternative, general traits of society when the returns of individual corrupt behavior depend – positively – on the prevalence of that behavior. This is the topic that we will discuss in Section 7.

After reviewing the theoretical literature, we turn to empirical issues related to corruption. The first of these relates once more to the illegal nature of corruption: it is difficult to obtain good data on corruption, and so measuring its incidence is challenging. In Section 8 we discuss different methodological approaches to deal with this problem. We also review the most relevant empirical evidence related to corruption using three different sources: cross-country, macro-evidence; results derived from field experiments; and evidence derived from lab experiments. Finally, in Section 9 we review the empirical research most closely related to our theoretical analysis: the individual incentives of bureaucrats to participate in corruption and the industrial organization aspects of bribing markets.



2 DELEGATION AND AGENCY

Agency theory provides the standard model of delegation in economics. It portrays the relationship of a principal and her agent. The principal draws up the terms of the relationship, specifying compensation and instructions for the agent who enjoys better information or ability, but also has its own goals. In the framework of the rules, the agent is vested with the power of taking decisions on behalf of the principal.

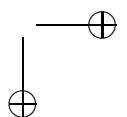
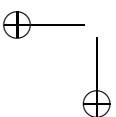
That framework is appropriate as a first step in describing the context of corruption: a principal – society, the public, or its representatives – drafts rules and entrusts the “power” to take certain decisions – issue a license, inspect and report on tax returns or emissions – in applying these rules to better informed officials. These decisions, on the other hand, affect third parties – citizens, entrepreneurs, regulated firms. Corruption is the abuse of entrusted power to bend the rules – take the wrong decisions – for private gain, most often as a result of illegal – voluntary or imposed – transactions with the affected third party.

Following the seminal paper by Tirole (1986), early works in agency theory model corruption indeed as a form of collusion between an official – bureaucrat – and an entrepreneur that the former is supposed to supervise.¹ A canonical model in this literature is Laffont and Tirole (1991), where an agency is in charge of regulating a firm with unknown costs. In the traditional regulation model, the principal – Congress – faces an information disadvantage with respect to the firm. The firm’s information relates to its costs, which may be high or low. The cost of any level of output may also be affected by private, unobservable effort – resources – that the firm puts into the activity. Effort imposes a private cost for the firm, $\Psi(e)$, which does not show in the accounting books. The principal designs rules that instruct the firm to produce a level of output q and also to put an amount of effort e into the activity, and the pair (q, e) could vary by “type” of firm, in exchange for some compensation that at least covers the firm’s total costs. The instruction must be incentive compatible. That is, the firm must have incentives to choose the level of output and effort that corresponds to its true cost function.

Let $\beta \in \{\underline{\beta}, \bar{\beta}\}$ be the two possible types of the firm, and $C(q, e; \beta) = (\beta - e)q$ the cost function for each of the types. The objective for the principal is to maximize the expected surplus for society, taking into account that public funds have a shadow cost. The asymmetry of information between the principal and the firm is modeled by assuming that β is the realization of a random variable that the agent – but not the principal – observes and takes the value $\underline{\beta}$ with probability v and $\bar{\beta}$ with probability $1 - v$. The asymmetry of information allows the firm to obtain information rents: a low cost firm can always claim that its costs are high, then produce the output \bar{q} and incur the accounting costs $(\bar{\beta} - \bar{e})$ expected from a high cost firm, but do so at a lower private cost: $\Psi(\bar{e} - (\bar{\beta} - \beta))$ instead of $\Psi(\bar{e})$. Thus, for the firm to have incentives to claim low costs when it does have low costs, it must expects profits (rents) of at least $\pi(\beta) = \Psi(\bar{e}) - \Psi(\bar{e} - (\bar{\beta} - \beta))$.

A specialized agency may improve matters for the principal by working close to the industry and so collecting industry-related information. Suppose that such agency may obtain (hard) evidence on the true value of β with probability μ . With that probability, and as long as the agency may be trusted to report truthfully, the principal won’t need to pay the extra profit

¹ Robert Klitgaard (1988) with his influential book about corruption contributed to popularize the principal–agent approach for describing and analyzing most models of corruption.



$\pi(\beta)$. But the agency may choose not to report truthfully, i.e., may hide information obtained. Corruption – capture – here would mean exactly that: an agreement – collusion – between the firm and the agency by which, in exchange for a bribe, the agency claims not to have obtained any evidence on β when it actually has.²

Laffont and Tirole investigate the optimal regulation scheme (effort and outcome) and agency compensation policy when the agency is corruptible. In order to prevent an agreement between agency and firm to hide the former's information, the principal must compensate the agency with a bonus s whenever it produces evidence of the firm's low costs. This is the extra cost that corruptibility imposes on the principal. Still, the use of an agency has its positive side: when information is obtained by the agent, the principal saves the information rents $\pi(\beta)$. The parameters of the optimal regulation and compensation scheme, s , \bar{e} , e , \bar{q} , and q balance the costs for the principal of distortions and rents.

This suggestive model may be taken as the starting point of our tour of the microeconomics analysis of corruption. We may take from it that even a corruptible agent may be of some value to the principal. The cost corruptibility imposes, the addition of a coalitional incentive constraint, must be compared to the cost linked to the “market failure” that it is supposed to address, a theme that later will be taken up by Acemoglu and Verdier (2000). However, this concise model abstracts from most of the interesting nuances related to corruption.

To begin with, the model concludes that it is always in the interest of the principal to make sure that corruption is prevented. But this “equivalence principle,” first stated in Tirole (1986), does not generalize. As Celik (2009) has shown, when the firm may have more than two types, incentive compatibility of agent and firm may not be separable and the “equivalence principle” may fail. Moreover, agents may be honest or dishonest, or have different propensities to participate in corruption transactions. Tirole’s “equivalence principle” is also not robust to situations in which the agent’s propensity for corruption is not common knowledge. Kofman and Lawarree (1996) and Strausz (1998) have shown that it may be optimal for the principal to allow some corruption when there is asymmetric information over the agent’s (corruptibility) type. Also, the model assumes bribery, as opposed to extortion: the agent can hide information, which is in the interest of the firm, but cannot produce false evidence against the firm, claiming, for instance, that the cost is low when it is in fact high. Khalil et al. (2010), in a model very similar to Tirole (1986) but with the possibility of forgery of (soft) evidence, show that when choosing among two evils, bribery and extortion, it may be worth accepting the former so as to prevent the latter. Note that bribes are always a cost to the firm, so that extortion imposes a cost on good behavior whereas bribery imposes a cost on bad behavior.³

The model also assumes “efficiency in collusion.” That is, with probability μ , agent and firm know each other’s information, so that their negotiations always lead to agreement when

² Interestingly, Kessler (2000) shows that collusion between agent and firm does not impose any cost on the principal if what the agent can monitor is the choice of effort, e – the moral hazard aspect – rather than the cost structure, β – the adverse selection parameter.

³ However, there are situations where bribing (capture) may be socially more costly than extortion. Auriol (2006) adapts Tirole’s model to a procurement setting in which due to transaction costs, the optimal policy is to organize an open tender only for large purchases. The agent has private information regarding the uncertain demand. If demand is high, and it is optimal to organize an open tender, the firm may bribe the agent to avoid it. Then, capture leads to an inefficient decision. If the demand is low, the agent can threaten the firm with organizing an open tender when it is not optimal to do so. In equilibrium, extortion only implies a redistribution of rents between the firm and the agent but the optimal policy is implemented.

there is room for it. But typically, information is not symmetric even between agents and firms, so that collusion is itself a problem. More importantly, the asymmetry of information, and so how difficult these agreements are, and then how corruptible the agent ends up being, may depend on the rules drafted by the principal. This research question has been pursued by Baliga and Sjostrom (1998) in a static setting and by Pechlivanos (2005) and Chassang and Padró-i-Miquel (2014) in a repeated game framework.

The difficulties of corrupt agreements not only come from asymmetries of information between agent and firm, but also from what is sometimes termed the “hold-up” problem: being typically illegal themselves, these agreements cannot be enforced by courts of law, and so the parties need to find ways to make them self-enforceable. The design of career paths for officials, penalties and rewards, and other instruments, affect the difficulty that agents and firms would find in reaching credible agreements (e.g., Lambert-Mogiliansky, Majumdar and Radner, 2008, Buccirossi and Spagnolo, 2006, Dufwenberg and Spagnolo, 2015). This is another point worth investigating.

Also, the model that we have discussed deals with giving the firm incentives – to exert effort, to reveal type – whereas the probability that the agent obtained information, μ , was assumed exogenous. But finding information may be a costly activity itself, so that the probability of success may depend on the agent’s effort. Giving the agent incentives may then be another issue in dealing with corruption. Here subtle questions may come to the fore (Mookherjee and Png, 1995). For instance, collecting a bribe may itself become a motivation for the agent to put effort into collecting information, and so there is a delicate balance between compensating officials, fighting bribery, and making regulation a success. Likewise, the selection of officials, when candidates may have differences in their cost – shame, for instance – of entering into illegal activities, may be an issue worth investigating (Besley and McLaren, 1993).

We will next discuss these and some other questions that pop up when we look to less abstract models of the relationship between principals, their agents, and affected third parties.

3 FINE DETAILS IN ILLEGAL NEGOTIATIONS

We start by reviewing some of the issues that agents and firms find when aiming at illegal agreements. As we have mentioned in the previous section, an illegal agreement (bribe in exchange for bending the rules) is always tainted by fragility. If a firm pays an official a bribe – say, in cash – in exchange for an illegal “favor,” and the official does not deliver, the firm may not enforce the “contract” by bringing it to a court of justice. Or vice versa: if the firm obtains the favor before the bribe is paid, the official cannot take the firm to court if the firm does not honor its promise. If the official – agent – and the firm interact repeatedly, the loss of future – illegal – benefits may be sufficient to prevent defection, just as with collusion in oligopolist markets. Lambert-Mogiliansky et al. (2008) study a model with these characteristics. An official – bureaucrat – in charge of issuing – on behalf of the government – licenses to operate in a market meets an entrepreneur who may qualify for the task and have a willingness to pay for the license, v , which is her private information. From the point of view of the official, this valuation is a draw from a random variable. The entrepreneur must first incur some cost $c > 0$ to prepare the application. But when the entrepreneur shows up at the official’s window, the official may decide to ask for a bribe b in order to issue the

license.⁴ Just as in the “lemons problem” (Akerlof, 1970), a one-shot game between entrepreneur and official will result in the entrepreneur never incurring the cost of application: if she expected a bribe b , then she would incur the cost if $v - c \geq b$. Thus, when the entrepreneur shows up at the window, the bureaucrat’s beliefs must be that the entrepreneur has a valuation $v \geq b + c$, and so he should ask for a bribe b' that is higher than the initial one, but we can repeat the argument for b' , and so on. This is the type of hold-up problem that prevents “trade” between the entrepreneur and the official. The outcome is not only very inefficient socially – if the entrepreneur does qualify for the license – but also for the parties involved. The bureaucrat would prefer committing to a bribe, but there is no enforceable contract that she may offer to that effect.

Repetition, but also reputation, is a way out of this dilemma. If a sequence of entrepreneurs visit the window and observe previous bribe demands, then bribery – and license issuing – may be an equilibrium. Each entrepreneur incurs the cost if they observe that the same – equilibrium – bribe, say b^* , has always been demanded in the past, provided their willingness to pay is $v \geq b^* + c$. But they never incur the cost of application if they observe that some other bribe has been demanded before. For a sufficiently low discount factor, in equilibrium indeed the official prefers to behave as expected by the entrepreneurs. The cost of losing their reputation, and then not receiving applications – and bribes – in the future is sufficient to restrain the official from “exploiting” an applicant who has shown up, implying a valuation $v \geq b^* + c$, by demanding a bribe of at least $b^* + c$.

Another, ancient way that parties may use to solve the hold-up problem is to exchange hostages. In the case of corruption, a piece of evidence of illegal payments or exchanges may serve as a “hostage”: if one of the parties does not deliver, the other may threaten to come forward with the evidence, which would penalize the deviant. Of course, the drawback of such a hostage is that evidence incriminating one party also incriminates the other. Threatening by shooting oneself in the foot may not be credible, which renders the hostage useless.

Related to this, Buccirossi and Spagnolo (2006) have pointed to subtleties that may turn well-intentioned policies designed to thwart bribery into weapons to make it possible. In particular, they discuss how leniency policies, whereby a reporting party may be exempted from penalties, may in fact make otherwise unfeasible bribery possible. Indeed, that policy makes the threat to come forward credible, and so makes it a useful hostage. Free from penalties, a party’s threat of reporting becomes credible, and then bribery may be sustained.

A particular form of leniency proposed in 2011 by Kaushik Basu became the center of a heated debate. With India and harassment bribes – extortion – in mind, Basu proposed to make bribe-giving not only legal, but also a way of recovering any bribe paid. The measure could be coupled with a double fine for bribe-taking officials (see Basu et al., 2014). There were several issues with the practicality of this measure, and the lack of trust of the police, i.e., of the “officials” in charge of receiving the report and acting upon it, was not the least of them. But even without this problem, Dufwenberg and Spagnolo (2015) have pointed to some other aspects that may render the policy counterproductive. Indeed, assume that the bureaucrat has to put some effort into studying the application for a license, for instance to check that the

⁴ Note that this is a model of extortion.

application meets the requirements.⁵ A bribe may then be the reward for that effort. Without Basu's policy, the entrepreneur would have to bribe the official, but then the license would be issued. On the other hand, if Basu's policy is implemented, and if the bribe is paid and the license issued, then the entrepreneur would have incentives to report the bribe and get it refunded. Anticipating this, the bureaucrat will not ask for and get a bribe, but will not spend the effort in issuing the license either. Corruption would be eradicated, but if the license is socially valuable, the remedy may not be all that positive.

On the contrary, if not issuing a license when it should be issued had a cost for the official – perhaps because lack of diligence may be detected and punished – things change drastically. Here again repetition of interaction is crucial. Suppose that entrepreneurs apply for licenses in sequence. Without Basu's policy, the bureaucrat may refuse to issue a license if a bribe is not offered, and issue it otherwise. Each entrepreneur would offer the bribe as long as the license had been issued anytime in the past that the bribe was offered, but would not offer it otherwise. As before, reputation sustains this as an equilibrium if the bureaucrat is sufficiently patient. The result would be that licenses would be issued, and bribes paid and not reported. But if Basu's policy was implemented, this equilibrium would no longer be possible: each entrepreneur would have an incentive to report the bribe once they had obtained the license. Under the policy, the bureaucrat would never ask for or accept a bribe, and yet she will issue the licenses since not doing so is now costly. Thus, licenses would still be issued and bribery eradicated.

On the same theme on the effect of intervention policies and their effect on the dealings between officials and entrepreneur, Chassang and Padró-i-Miquel (2014) have recently studied the efficacy of “whistle-blowing” and the limits of it. Here the issue, rather than bribery, is how well the entrepreneur who is not complying with regulation may infer that the official – or some employee – has reported that non-compliance. By introducing noise into this inference, for example by not always intervening in a report by a whistle-blower, the principal may complicate that inference, making it more difficult for the entrepreneur to retaliate, which makes it safer for the whistle-blower to report.⁶

4 SIZE, SELECTION, AND COMPENSATION OF THE BUREAUCRACY

As we have mentioned in Section 2, the costs of bureaucratic corruption must be weighed against the market failure that the bureaucracy is designed to correct. Against the view, sometimes defended vehemently (e.g., Shleifer and Vishny, 1994)⁷ that regulation and public intervention only extract rents and impose inefficiencies, other scholars have analyzed the trade-off between the inefficiencies associated with regulation and the costs from market

⁵ Dechanaux and Samuel (2012) analyze a model where the inspector must exert effort in order to find hard evidence, and the entrepreneur can either pre-empt inspection or offer bribes ex post. They also consider the hold-up problem that we have discussed and repetition as a way to enforce illegal dealings.

⁶ Pechlivanos (2005) also analyzes a principal–agent–firm relationship where the repeated interaction is the mechanism for enforcing corrupt contracts. The principal may decide to implement the agent's decision or to review it. Most interestingly, it is optimal for the principal to make the auditing not observable by the firm. By doing so, he introduces noise into the agent–firm relationship, making the corruption transaction less likely.

⁷ An influential book by De Soto (1989) analyzing the costs imposed by bad regulation in Peru has been widely used to support this view, although De Soto himself was discussing regulatory – formal sector – reform.

failure that motivates that regulation. Regulatory capture or bribery is one of the potential costs of intervention. Acemoglu and Verdier (2000) take this approach and model the trade-off between allocating resources (human capital) to regulation and allocating them to entrepreneurship. They assume a fixed amount of potential entrepreneurs who make their career choice between productive activities and bureaucracy. The reward of the latter is a wage, w , set by the government, whereas the payoff for the former are the profits in the market. These may be obtained using a “good technology,” privately more expensive but “clean,” or a “bad technology,” cheaper but imposing a negative externality on society.⁸ The government regulates the market by imposing taxes, τ , on the bad technology – and/or subsidies for the good one. However, implementing this policy requires inspection, and so requires a bureaucracy. The larger the size of the bureaucracy, the larger the number of “firms” that can be inspected: inspecting one firm requires hiring one bureaucrat. Taxes – and subsidies – and wages determine the expected profitability of each occupation, and so the career choices of individuals. A certain size of the bureaucracy is necessary for the probability of inspection to be large enough so that entrepreneurs choose the good technology. On the other hand, the larger the bureaucracy the larger the resources diverted from production.

The authors then assume that bureaucrats are corruptible – and bribery is detected with an exogenous probability, θ , resulting in the bureaucrat been fired. In fact, bureaucrats, if corrupt, become bounty-hunters: they simply ask each entrepreneur they inspect for a share, α , of the cost they could impose with a bad report, and this irrespective of the good technology; that is, the tax, τ . Thus, a corrupt bureaucracy is a total waste: whether inspected or not, entrepreneurs do – privately – better by choosing the bad technology. Therefore, the real choices for government are either *laissez faire* (no bureaucracy) or a bureaucracy without incentives to be corrupt. This requires a wage $w \geq \frac{1-\theta}{\theta} \alpha \tau$.

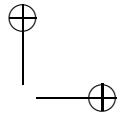
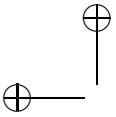
A bureaucracy, one that is also corruptible, imposes a cost, but this cost makes it possible to fix a market failure: an externality that would otherwise result in the choice of the wrong technology. In fact, corruptibility calls for a larger bureaucracy: the larger the probability of inspection the lower the taxes, τ , needed to induce entrepreneurs to choose the good technology – a tax that is not collected if regulation succeeds – and so the lower the wage needed to prevent corruption, w .

Note that the wage w must be above the reservation utility of the bureaucrat. That is, corruptibility demands an “efficiency wage.” For the threat of dismissal to be sufficient to prevent corruption, the wage in the bureaucracy must be larger than in alternative occupations. Thus, large bureaucracies with high wages may not be sufficient evidence to conclude that a bureaucracy is always pure waste, but a necessary cost for reaping the benefits of regulation.

The trade-off between the cost of inducing honest behavior by the bureaucrat and its benefits is also the topic of a study by Besley and McLaren (1993). They take the size of the bureaucracy and also the proportion, v , of entrepreneurs or citizens abiding by the regulation as given.⁹ Suppose the tax, τ , is exogenous, perhaps set by the legislative. The authors consider the possibility that potential bureaucrats are heterogeneous in type. In particular, each hired bureaucrat may be dishonest with some probability, γ , just as they were in Acemoglu and Verdier (2000). But with probability $1 - \gamma$ she is honest and would never take a bribe. Also, a bribe payment is detected with probability θ .

⁸ Formally, the authors assume that it is the good technology that generates an externality, but a positive one.

⁹ The authors present their model in terms of tax-compliance inspection, so that what is taken as fixed is the proportion of tax-payers who owe a tax τ .



In the face of this, the government (principal) has three possible policies when setting the compensation w : pay efficiency wages, which will discourage even dishonest bureaucrats from taking bribes; pay reservation wages, which will attract both types of potential bureaucrats but does not prevent dishonest ones from taking bribes; and pay what the authors call capitulation wages, that is, wages below the reservation wage, which will attract only dishonest bureaucrats, lured into the bureaucracy by the prospect of obtaining additional, illegal income $\alpha\tau$. The authors embed this framework in a dynamic model where some bureaucrats retire in each period for exogenous reasons, and a portion θ of illegal dealings is detected. The first policy is the most expensive, but results in a bureaucracy that always “acts” honestly. The second policy is less expensive but some bribery occurs in every period. Dishonest bureaucrats are caught and fired, and are replaced by average bureaucrats, so that with time the bureaucracy improves. Finally, capitulation wages only attract dishonest potential bureaucrats, so even if caught taking bribes, there is no reason to fire a bureaucrat. Such bureaucracy is the least expensive, but corruption is totally widespread.¹⁰

Note that, although the authors do not explore this issue, a highly paid bureaucracy would be associated with a higher cost of not complying with regulation (if inspected, τ) for the entrepreneur, whereas lower wages in the bureaucracy would be associated with lower costs of non-compliance: if inspected, τ if the bureaucrat is honest – i.e., never with capitulation wages – but only $\alpha\tau$ otherwise. Thus, if the level of non-compliance was the entrepreneur’s decision, a less expensive bureaucracy would be a less powerful instrument to correct the market failure.

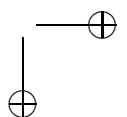
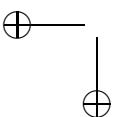
Mookherjee and Png (1995) allow for this decision on the part of the entrepreneur. For instance, we may let the social harm imposed by a technology, x – say, pollution – to be a continuous variable, so that the tax owed by an entrepreneur who chooses a value of x is τx . We may also assume that finding out evidence of this harm is costly for the bureaucrat, so that μ is a function of – private – effort e , and effort has a cost of $\Psi(e)$. The bureaucrat is expected to report the value of x , and in that case gets a compensation of wx . However, even if she finds evidence, she may decide not to report and solicit a bribe, a transaction that is exogenously detected with probability θ and, if detected, results in a penalty $p_b x$ for the bureaucrat. Both τ and p_b are chosen by the legislative, and so are not the government’s choices. The government can only choose the bureaucrat’s compensation w .

Suppose that

$$\tau(1 - \theta) > w + \theta p_b. \quad (16.1)$$

The left-hand side is the entrepreneur’s gain per unit of x of not been reported, and the right-hand side is the bureaucrat’s cost per unit of x not reported, having obtained evidence. If (16.1) holds, then even if evidence is obtained, the bureaucrat and the entrepreneur can mutually agree on a bribe so that the bureaucrat does not report when she finds evidence. We may assume that they equally share the proceeds in this case. If (16.1) is violated, then corruption is prevented, but otherwise the bureaucrat never reports. The entrepreneur and the bureaucrat play a simultaneous-move game when choosing x and e respectively, although the payoffs of the game are different depending on whether (16.1) holds or not.

¹⁰ The authors discuss what is the best policy for the government and argue that efficiency wages (high compensation and high honesty) is best for rich countries, but one of the other two may be best for less developed realities.



It may be easily shown that, for a total surplus maximizing government, the optimal policy is to set w so as to prevent corruption – (16.1) is reversed. Then, the payoff for the bureaucrat is $\mu(e)wx - \Psi(e)$. Therefore, if Ψ is convex, effort is increasing in w , and as a consequence the equilibrium choice of x is decreasing in w , since the probability of detection is increasing in w .

Note that the government has two objectives when choosing w : give the bureaucrat incentives for diligence, and also prevent corruption. (Also note that, as opposed to what was the case in Section 2, this model deals with providing incentives to the “inspector,” not to the agent.) The prospect of soliciting a bribe from the entrepreneur is an alternative motivation for the bureaucrat to put effort into finding evidence. However, when distinguishing between bribery-induced and indolence-induced underreporting is possible, bribery is never an efficient way to provide incentives.

Things are different if what the government detects with probability θ is the size of the harm x when it is not reported, but distinguishing when this is due to bribery or to the bureaucrat having failed to find evidence is not possible. In that case, having two objectives and only one instrument, and under some conditions (e.g., the marginal cost of effort not increasing greatly), totally avoiding corruption may be too costly and then suboptimal.

5 COMPETITION

Competition, that is, the “industry organization,” is a recurrent topic in the literature on corruption, whether it refers to the “supply side” – bureaucrats or officials – or the “demand side” – entrepreneurs or firms. In fact, the latter is one of the “usual suspects” when it comes to explaining the origin of corruption. Rents, either from market power or from natural resources, would be the booty on which corruption would prey, and so the ultimate cause of corruption. (Ades and Di Tella, 1999, were perhaps the first to document the relationship between rents and corruption.) Taking this point a little further, competitive markets should then be free from the threat of corruption.

The problem with this view is one of causality and also one of implication. Should we see the number of firms in a market, or equivalently, price-taking behavior, as a sign of the absence of corruption? Is the existence of rents a prerequisite for corruption to exist? Bliss and Di Tella (1997) argue the fallacy behind these conclusions. Consider an industry of small firms $i \in [0, 1]$ with heterogeneous entry (sunk) costs k_i who need to obtain a license from a corrupt official. The official does not observe the entry cost of applicants for the license, and can set a bribe b – on top of, say, a legal entry fee r that perhaps is motivated by an externality. If firms have constant and common marginal costs, they will enter only if $k_i \leq b + r$. Thus, only $F(b + r)$ firms would enter, and the official would appropriate $bF(b + r)$ bribe revenue. Optimally, the official would set a positive b^* , that solves

$$\max_b bF(b + r). \quad (16.2)$$

Corruption introduces a wedge between average cost and price, and the bribe allows the official to appropriate the difference. That is, corruption generates its own rents even though firms are price-takers.¹¹

¹¹ Bliss and Di Tella (1997) distinguish “deep competition” from competition: if the profits upon entry depend on some efficiency parameter, then the higher the efficiency the higher the profits upon entry that firms can make and the

This model is consistent with the already mentioned view that any – or, at least, most – regulation is simply a way for a class of politicians or officials to extract rents. Absent any reason to limit entry, corruption would be viewed as – one of – the instruments to cash those rents.

In a recent paper, Amir and Burr (2015), building on similar arguments, put the emphasis on how corrupt officials have a vested interest in shaping the structure of the industry so as to also affect behavior. That is, if firms are not price-takers, by limiting entry the official may not only affect the number of firms, but also the margins over prices and so firms' profitability. In fact, if firms are not small, rents are maximum when only one firm operates, and so the best that a corrupt official can do is to issue only one license and then extract – a fraction of – monopoly profits.

The authors also consider the case of a pre-existing industry with some already licensed firms, or the presence of some firms that have opted to operate in the “shadow economy” – i.e., without a license.¹² The official will limit entry, but under quantity competition this time she may allow for entry of more than one firm. From a second-best point of view, and depending on the existing number of firms, total entry may be too large or too small.

Amir and Burr (2015) also investigate the other side of competition, i.e., competition between officials. This supply-side aspect of competition has been present in the theoretical and practical debate on corruption at least since Rose-Ackerman suggested that it may be an instrument for corruption control (see Rose-Ackerman, 1978). Continuing with the example of license issuing, there are two ways – at least – in which officials may “compete.” First, more than one official may be authorized to issue the same license. Second, more than one license, issued by different officials, may be required to operate. Note that in this second case the “goods” sold by the different officials are complements. That is, officials are in a vertical relationship, and do not exactly – horizontally – compete.

Shleifer and Vishny (1993) consider this second case. As in any problem of complementary goods, each official imposes a vertical externality on the other when deciding how high a bribe to solicit from entrepreneurs. The externality results in double-marginalization, which implies bribes in excess of what is bribe revenue maximizing. Thus, as in the case of vertically related producers, centralizing the issuing of licenses would result in higher rents for the official(s), but also a lower burden for the entrepreneur and so higher entry and welfare.¹³

Shleifer and Vishny assume simultaneous moves by all officials in charge of issuing licenses. If, on the contrary, licenses must be obtained sequentially, the same type of vertical

official can extract through bribes. The effect of this “deep competition” on the rents that the official can appropriate has an ambiguous sign.

¹² Choi and Thum (2005) explore the relationship between the “shadow economy” and the ability of corrupt officials to distort the working of the market. Instead of assuming the selection between the official and the shadow economy as exogenous, as Amir and Burr (2015) do, they consider explicitly the entrepreneur’s choice between applying for a license and operating in the underground economy. They show that this second option severely limits the official’s ability to extract bribes and mitigates the distortions it creates.

¹³ Celentani and Ganuza (2002) explore a different type of externality and obtain a similar result. Consider a principal that delegates to a group of agents (bureaucrats) the contracting of some service with a group of providers. The principal sets the level of quality/production and he is aware that a bureaucrat can allow a provider to supply a lower level of quality in exchange for a bribe. Bribes and corruption profits are increasing in the quality required. Taking as given penalties and monitoring probabilities, the principal’s best response is to set lower levels of required quality when he expects a higher level of corruption. If agents take decisions individually and do not internalize the externality, the number of corrupt transactions is high but profits for corruption (and welfare) are low. On the other hand, if bureaucrats take the decision jointly, they would maximize their profits by reducing the number of corrupt transactions, increasing also total welfare.

externality results in the same type of result: solicited bribes are too high with respect to the revenue-maximizing levels.¹⁴ That is, vertical competition is not a good prescription to mitigate corruption, and a “single window” for obtaining all the required licenses may be advocated based on, among other benefits, its effect on bribes.

Renewal of licenses may be regarded as nothing but a particular case where entrepreneurs need several licenses and apply for them in sequence. However, it introduces novel nuances, in particular a credibility problem for the official reminiscent of the one faced by the durable-goods monopolist. As in Choi and Thum (2003), suppose the entrepreneur needs to apply for a license in each of two periods, and as in Bliss and Di Tella (1997), suppose the official does not know the entrepreneur’s – per period – fixed cost, k_i , when they first meet.¹⁵ If the official could commit to the bribe she will solicit in each period, she will choose b^* that solves (16.2). Entrepreneurs with fixed costs below $b^* + r$ would apply in the first period, and the rest would never apply. But if the official can’t commit to honor her promises, she would be unable to commit not to solicit in the second period a lower bribe, b_2^n , to firms that did not apply in the first period, so

$$\max_{b_2^n} [F(b^*) - F(b_2^n)] b_2^n.$$

Expecting this, an entrepreneur with a fixed cost k_i above, but close to, $b^* + r$ will not apply in the first period.

Consequently, charging b^* in both periods without price discriminating between new and old applicants is not an equilibrium if the official has no means to commit. In equilibrium, the official will in fact solicit a bribe b_1 , which will induce less entry in the first period – entrants have fixed costs strictly lower than $b^* + r$ – and will price-discriminate entrepreneurs in the second period, depending on whether they are applying for renewal or are applying for the first time. The bribe solicited to the latter will be lower than b^* , so that in the second period entry will be larger than with commitment.

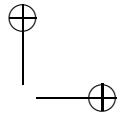
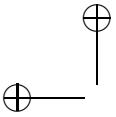
Choi and Thum (2003) use this model to investigate one often-suggested measure for curbing corruption: job rotation.¹⁶ Replacing the official in charge of a particular task can have positive effects for different reasons, like making reputation building more difficult and revelation of (corrupt) types. In the present context, job rotation makes price discrimination more difficult. The authors compare the results obtained when one official is in charge of issuing the license in both periods with an alternative model where the official is replaced – with some probability – by a new official in the second period. This unambiguously results in more entry, and so higher welfare, if the new official does not observe who obtained a license in the first period. But if renewal is distinguished from first-time application, the results are ambiguous.

But let us return to “horizontal competition” among officials. Amir and Burr (2015) obtain that the horizontal externality between officials will increase entry, so that, optimally, granting the right to issue licenses to more than one official increases entry and welfare. Although the authors do not analyze price – bribe size – effects of competition, this is probably the sort

¹⁴ However, Lambert-Mogiliansky, Majumdar and Radner (2007, 2008) show that reputation or repetition of the application process may result in the opposite.

¹⁵ Choi and Thum (2003) model the asymmetry of information in terms of the value of the license, v , rather than the fixed costs.

¹⁶ See, for instance, Transparency International (2006).



of result that is behind the idea that competition among officials may help control bribery, as argued by Rose-Ackerman (1978). Drugov (2010), one of the first attempts to formally analyze this issue, points to more subtle effects of competition. He considers its effect on the price – i.e., on entrepreneurs' outside option when bargaining with corrupt officials – and also on the incentives for entrepreneurs to invest in compliance, as in the model by Mookherjee and Png (1995). Entrepreneurs may have different costs of compliance, increasing in the negative externality they create if they do not comply. Officials may be honest or dishonest.

With a monopolistic official, if a firm does not incur the cost of qualification – compliance – then she may risk facing an honest official who would not issue a license. If, on the other hand, the official happens to be dishonest, the entrepreneur would obtain the license but would be asked to pay a bribe, both whether qualified or not. Thus, the trade-off is between saving in expected costs of compliance and risking not obtaining a license.

If many officials can issue the license, unqualified entrepreneurs who meet an honest official do not get a license, but then can reapply – at a cost of delay, perhaps – to another official and this time be luckier. Eventually, she will get a license. Thus, taking the decision to invest in compliance as given, the monopoly results in less licenses being issued to unqualified entrepreneurs, and so less social harm.

However, the outside option of qualified entrepreneurs is larger under competition. Indeed, a qualified entrepreneur who has the misfortune of meeting a dishonest official will have to pay a bribe (extortion) if she is to get a license. Her outside option is to give up the profits she can make in the market. On the other hand, under competition, the entrepreneur can try her luck with a new official, hoping to find an honest one. Thus, competition among officials increases the entrepreneur's outside option, so that the expected bribe is lower, if the bribe is endogenously set by negotiations.¹⁷ Thus, ex post, the monopoly regime results in – larger bribes but – ex post better allocations, whereas the competition provides stronger incentives to compliance.

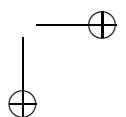
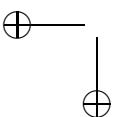
The papers offer one more example of the subtleties behind proposed reforms and the relationship between bribes and compliance, one that links to the observation by Khalil et al. (2010) that extortion is a bigger problem than bribery. For suppose that bribery, but not extortion, is the problem. Since monopoly increases the bargaining power of dishonest officials, and so the size of the bribe, monopoly unambiguously results in more compliance and lower harm.

6 BRIBE COMPETITION AND BACK TO DESIGN

In the previous section we have used the metaphor of a market (or a vertically related structure) to describe corrupt transactions. Potential suppliers of a essential input (license, for instance) trade with potential producers that use that input. We have discussed competition by the upstream suppliers and competition downstream. We may extend the metaphor to discuss competition among upstream buyers, the best example of which is public procurement.

Public procurement is managed by public officials who are instructed to follow very detailed rules to select both the terms of the contract and the identity of the contractor. Also,

¹⁷ Drugov (2010) analyzes a richer model, where an unqualified entrepreneur who meets an honest official, whether under competition or monopoly, can invest in qualifying and then reapply at a cost, say of discounting.



public procurement has often been found to be tainted by widespread corruption. Documented cases abound of officials – and relevant personalities with influence over officials – taking bribes to bend the rules so as to favor one supplier or to improve the terms of the transaction for her. Bribes have also been instruments to secure collusion among suppliers.

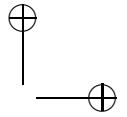
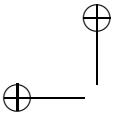
The theoretical literature has investigated several ways in which corruption may take place in procurement auctions. One is favoritism: the official in charge – auctioneer – may reveal the – honest – winning bid to some favored, pre-contacted bidder (Arozamena and Weinschelbaum, 2009, Burguet and Perry, 2007, Koc and Neilson, 2008), or some bidder she selects *ex post* (Lengwiler and Wolfstetter, 2010, Menezes and Monteiro, 2006). The favored bidder can then change her bid to match a rival's better bid or to improve her own winning bid.

This type of “favoritism” is also analyzed in Burguet and Che (2004), and Compte, Lambert-Mogiliansky and Verdier (2005), but with the addition of bribe competition. That is, the official bends the rules in favor of one supplier, but who that supplier may be is determined by bribe competition among all of them. At first glance, bribe competition may be considered just another form of price competition, so that the effect of bribery would be an increase in the price that the principal – government – pays for the supplies: the supplier who is in a better position to bid lower – a low cost supplier – is also in a better position to offer a larger bribe.¹⁸

Unfortunately, the effect of corruption is much more than this. Suppose, as Compte et al. (2005), that n *ex ante* symmetric suppliers compete for a contract in a first-price auction – the rules: they simultaneously quote a price offer, and the lowest offer is accepted. Each supplier, i , has a cost of delivering, c_i , which is her private information. Absent corruption, the supplier with lowest cost indeed bids lowest and so is selected as the contractor. But now assume that, after receiving all the bids, and before making them public, the official informs every supplier what the winning bid is, and allows them to make a bribe offer. Whoever makes the highest bribe offer, is allowed to submit a new bid to match the standing winning bid. If this is common knowledge at the time of first submitting bids, then no supplier has any incentive to submit a bid lower than the maximum acceptable one, P . Indeed, doing so can only reduce the price that she will get if she wins the ensuing bribe competition. True, this may discourage a competitor from offering a bribe, but it is always better to beat that potential competitor at the bribe competition without reducing the prize they will be fighting for at that stage. That is, bribery induces endogenous collusion at the highest acceptable price and, if there is no limit to the bribe that suppliers can offer, then the official appropriates all the extra rents: suppliers expect the same rents as without bribery. That is, in particular, the lowest cost supplier will win the contract, and bribery simply results in a higher price, P .

However, bribery has a second effect, if we assume that, for whatever the reason, there is a maximum bribe that suppliers can offer, B . That limit to bribe competition then results in a distortion in the allocation of the contract: the lowest cost supplier does not have the ability to secure the contract. In fact, if the difference between P and B is sufficiently large, then all

¹⁸ This argument was fully developed by Lien (1986) and Beck and Maher (1986), who conclude that as the contract goes to the most efficient bidder, corruption may not generate efficiency losses. Lui (1986) goes beyond that, and builds a model of queues in which bribing may generate higher total surplus. He shows that there may exist an equilibrium in which the size of the bribe positively depends on the client's opportunity costs of time. This equilibrium in which bribes determine the position in the queue, minimizes the waiting costs and improves efficiency. However, this is not a robust result: bribing may also reduce the speed of the queue.



contractors will be willing to offer the highest bribe B , and so the contract will end up assigned to them in a totally random way. Thus, bribery imposes an efficiency cost on top of a transfer of rents from principal to official.

In fact, efficiency costs should be expected even in the absence of limits to bribes. Suppose, as in Burguet and Che (2004), that suppliers' bids are evaluated, paying attention not only to the price but also to the quality offered. That is, each supplier i selects a price p_i and a quality q_i , and bids are evaluated according to some scoring rule that assigns to each pair (p, q) a score $S(p, q)$. To simplify, suppose, the scoring rule assigns scores according to the social value of quality, and this is $S(p, q) = q - p$. However, quality must be assessed by an expert, the official. Suppose that, when suppliers submit their bids of quality and price, they also submit a bribe offer, b_i . The official can manipulate the quality assessment, and so declare a true quality offer of q to be of quality $q + m$. She will take bids (p_i, q_i) and bribe offers b_i from all suppliers, and accept the highest bribe b_i submitted together with a bid (p_i, q_i) that can be "assessed as winning." That is, so that $S(p_i, q_i + m)$ is larger than $S(p_j, q_j)$ for all $j \neq i$. Consider only two suppliers, and suppose that each has a cost function for quality delivered, $C_i(q)$, with $C_1(q) \leq C_2(q)$ for all q . That is, supplier 1 is more efficient. Absent bribery, supplier 1 will outbid supplier 2 and also offer to deliver what is efficient: $q_1 = \arg \max_q q - C_1(q)$. She will set the price p_1 so as to obtain a score $q_1 - p_1$ that makes it impossible for supplier 2 to attain the same score with non-negative profits.

Bribery changes this. Supplier 1 would have to reduce the price by m in order to still be able to guarantee a win without bribing. That may be the best course of action, if m is small. Then corruption would in fact result in the same quality and a lower price for the principal.

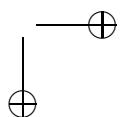
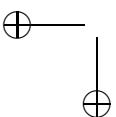
But that may turn out to be too expensive, unless m is really small. On the contrary, supplier 1 will find it in her interest to share the market with supplier 2, if m is larger. Both supplier 1 and supplier 2 will offer bribes, and both will win with positive probability.¹⁹ On average, the price will increase by more than the expected bribe, and the efficient firm will not be selected for sure, an inefficiency beyond rent transfers. Supplier 2 will profit from corruption, but even supplier 1 may see her expected profits increase. That is, bribe competition is not equivalent, but on the contrary, blunts price competition and serves again as an instrument for tacit collusion among suppliers.

But once we find ourselves discussing procurement models, there is a natural question that will take us back to the beginning of this journey: facing the threat of bribery, what rules should the principal instruct the – corruptible – agent to implement? For example, are first-price auctions, or the "true" scoring rule $S(p, q) = q - p$ best, once we know that the official may manipulate their outcome? Both Compte et al. (2005) and Burguet and Che (2004) provide *partial* remedies in the contexts they analyze. Thus, Compte et al. (2005) show that handicapping the – ex ante – most efficient supplier may destabilize the collusive outcome, and Burguet and Che (2004) show that de-emphasizing quality in the scoring rule reduces the effect of bribery.²⁰

But instead of partial remedies, we may be interested in finding out what the optimal designs are for procurement under the threat of bribery. This question is very much open as of now. It is a demanding one, particularly if the official can – illegally – negotiate with

¹⁹ Under complete information, the equilibrium will always be in mixed strategies.

²⁰ Other studies have discussed other measures that mitigate the incidence of corruption. A recent example is the paper by Auriol and Soreide (2015) that studies the consequences of debarment as a tool to deter corruption, and its consequences for collusion.



suppliers between the time the rules are designed by the principal and the time of selecting the contractor in – allegedly – application of the rules.²¹ We have answers for the other two possible cases.

Celentani and Ganuza (2002) analyze a procurement setting in which, as before, quality and price enter the preferences of the principal, $V(q) - p$; n potential bidders have private information about their production costs and the delivered quality q is certified by an official who can manipulate his report in exchange for a bribe. Then the principal designs the optimal procurement mechanism, anticipating that she will pay for quality as if no corruption exists, but will obtain an exogenous low level of quality q_C with some probability γ (the probability of corruption). They show that in order to implement the optimal mechanism, the principal must use a scoring rule $S(p, q, \lambda)$ in which the higher the probability of corruption, γ , the lower the official discretion λ (the weight of the principal's preferences for quality in the scoring rule). Moreover, Celentani and Ganuza (2002) assume a particular way in which corruption takes place and endogenize the probability of corruption γ . The official decides to become corrupt if he anticipates that his discretion λ will be high enough and then it is worthy to do so. The higher is λ , the higher the expected quality of the procurement process q and the higher the profitability of corruption (for replacing q by q_C). Then, the principal and the official play a simultaneous game, in which their best responses ($\lambda(\gamma)$) the optimal discretion level is decreasing in the expected corruption and the $\gamma(\lambda)$ corruption level is increasing in the expected discretion) determine the equilibrium level of corruption γ^* . Celentani and Ganuza (2002) undertake several comparative static exercises and show that contrary to conventional wisdom, corruption may well be increasing in competition. The intuition is that higher market competition (larger n) decreases the expected cost of quality, which may lead to higher expected quality being supplied, and higher incentives for the agent to become corrupt. If we assume that the agent has to incur some idiosyncratic costs when verifying the delivered quantity, q , a similar argument can be made regarding competition in the market for procurement officials. Higher competition allows the selection of a more efficient official, then the opportunity cost of reducing his discretion is higher. Since higher discretion implies higher profitability of corruption, a higher level of corruption may arise in equilibrium.²²

We also have some answers when the official and the supplier meet and negotiate only after the contract has been assigned and its terms determined. Suppose again that the suppliers' cost function – type – is the suppliers' private information. Also, as in Burguet (2014), suppose that after price p and quality q have been committed to by the contractor, the principal employs an official – inspector – to certify that the delivered quality is as contracted. Suppose that this

²¹ The standard approach to answering questions like this is to invoke the “revelation principle” and then focus on mechanisms – rules – that consist of simply asking each agent involved to reveal her type, making sure they have incentives to answer with the truth. This is because the principal is supposed to control and commit to applying any rules she designs. When an official can bend the rules, this commitment power is absent: the principal is only partially the designer. Not even the “revelation principle” can be generally invoked.

²² Laffont and N'Guessan (1999) obtain a similar result in a model of regulation similar to Laffont and Tirole (1991) in which the regulator optimally chooses the contract to be offered to the regulated firm and to the agent. They consider that agents may have different propensities for being corrupt. Then, the regulator may choose between eliminating corruption by providing incentive payments for good behavior to all agents or to save part of such incentive payments, deterring corruption only from the less corruptible agents. Laffont and N'Guessan show that greater competition among agents (better monitoring technology) may make this latter corrupt regime more attractive. The idea is that in the corrupt regime, the regulator has to distort agent incentives to reduce informational rents, and with greater competition these rents are lower.

inspector may be bribed. Several models of how bribe negotiations proceed can be posited, but for illustration suppose that the inspector is willing to certify any level of quality, irrespective of the real quality, in exchange for a fixed bribe b . In this case, it is in the principal's interest to prevent bribery, and the way to do so is to distort quality both at the bottom and at the top. That is, optimally, the principal gives up inducing quality from suppliers with low efficiency. This is a typical result in adverse selection models, as it reduces information rents for high efficiency suppliers. However, in this case this distortion serves to guarantee that low efficiency suppliers do not profit from claiming to have high efficiency, then committing to a high quality only to bribe the official and deliver minimum quality instead.

But optimally, the principal curtails quality also at the top end of efficiency types. Asking for high levels of quality when the supplier is highly efficient requires compensating her with higher prices. The principal may be willing to pay this higher price. However, a higher price also means a stronger incentive for low types to again claim high efficiency, commit to high quality, but then bribe the inspector and deliver the lowest quality.

Thus, that bribery imposes an efficiency cost, and in particular erodes the quality of delivered goods and services, seems quite a robust conclusion. This is so even once we take into account how the principal's best instruments respond to the threat of corruption.

7 MULTIPLICITY OF EQUILIBRIA AND THE PERSISTENCE OF CORRUPTION

In previous sections, we have presented corruption models that deliver a unique equilibrium, and organization designs and incentive systems that determine when agents are corrupt. While these models help in understanding important comparative statics and to design better institutions, they are unable to explain why societies with similar institutions and incentive systems have very different corruption levels. This is the motivation of a branch of the literature that explores multiplicity of equilibria in models with corruption. As Andvig and Moene (1990) point out, "corruption corrupts." The expected profitability of being involved in corrupt activities depends on the number of other people doing so. For example, the "moral cost" or guilt associated with corruption may decrease as the number of corrupt people increases. By the same token, even if a social norm exists that stigmatizes corruption, the corresponding loss of reputation is likely to decrease when corruption is widespread. Additionally, if the resources devoted to monitoring corrupt activities are limited, the probability of detection may decrease when many others are corrupt. Andvig and Moene (1990) show that when some of these forces are in place, multiplicity of equilibria arises. In particular, we can find a low corruption equilibrium in which the cost of corruption is high and only individuals very prone to be corrupted are involved, and a high corruption equilibrium in which due to the effects described above the costs of being corrupt are relatively low, and many more individuals decide to be corrupt.

We can formalize these ideas with a very simple monitoring model based on Cadot (1987). Consider, as before, that an official is in charge of issuing licenses to operate in a market. The official may decide to extort the potential entrepreneur by asking for a bribe b . The incentive system is as follows. The official receives a wage w , and with probability θ is monitored by another official (inspector). The inspector can also be honest or corrupt. If the inspector is honest, the corrupt official is fired and loses his wage. If the inspector is corrupt,

he takes the bribe b and does not report anything. The official has to incur an idiosyncratic, moral cost when he asks for a bribe, c_m , distributed according to a cumulative distribution function (c.d.f.), G . For simplicity, we assume that γ is the proportion of both corrupt officials and corrupt inspectors. Then, an official becomes corrupt if the following condition holds:

$$(1 - \theta)(w + b) + \theta\gamma w - c_m > w \iff (1 - \theta)b - (1 - \gamma)\theta w > c_m.$$

The level of corruption is given by the probability that the moral cost of the official is below the threshold $c^* = (1 - \theta)b - (1 - \gamma)\theta w$. Then $\gamma^* = G((1 - \theta)b - (1 - \gamma)\theta w)$ defines the endogenous probability γ^* that the official/inspector are corrupt. Notice that, as both sides of the equality are increasing in the level of corruption, multiplicity of equilibria may arise. We can have a low level of corruption, where most of the inspectors are honest and the expected penalty for corruption is large. We may also have a high level of corruption, where the probability of being matched with a corrupt inspector is high, and consequently the expected penalty (the probability of losing the wage) is lower. The main implication of the multiplicity of equilibria is that only big anti-corruption campaigns that are able to move from the high corruption equilibrium to the low corruption equilibrium, are really effective when corruption is widespread.

Tirole (1986) proposes an alternative model of collective reputation with multiple equilibria and discusses the problem of persistence. He considers a dynamic setting in which a population of principals are matched with a population of agents every period. Each principal assigns one task to the agent with whom he is matched. There are two tasks, 1 and 2. Task 1 generates more surplus and higher profits for the principal if the agent is honest, but yields lower profits if the agent is corrupt. The agent can be of three different types, corrupt (p_c), honest (p_h) or strategic (p_s), where $p_c + p_h + p_s = 1$. The behavior of the first two types is fixed, but the actions taken by the strategic type depend on his incentives. The strategic type receives a higher payoff when he is assigned the efficient task, task 1, that is, $w_1 > w_2$. Also, independently of the task, he can obtain an additional benefit from being corrupt, b , which does not depend on the task. Then, he may give up corrupt gains if that sufficiently increases the probability of being assigned task 1 in the future. The matching also has several other properties: at the end of each period the agent either (a) dies and is replaced by a new agent with some exogenous probability, or (b) is matched with another principal. Finally, principals have only imperfect information about the past behavior of agents. They observe whether the agents have been involved in corruption in the past with noise. A bad realization of the binary signal is more likely when the number of periods that the agent has been corrupt is larger (at a decreasing rate). Tirole provides several interesting results. First, this economy may have multiple equilibria. One is a low corruption equilibrium in which strategic agents are not involved in corruption. On the one hand, principals assign task 1 to agents having a zero record of corruption since the collective reputation of agents is good: *ex ante*, agents are honest with a probability $p_h + p_s$. On the other hand, strategic agents choose to be honest in order to preserve their individual reputation. A high corruption equilibrium may also arise. In this equilibrium, strategic types are involved in corruption. Collective reputation of agents is bad, the aggregate probability of corruption, $p_c + p_s$, is also large, and so even a clean individual reputation (no evidence of corruption in the past) is not enough to convince the

principal to assign the agent task 1. Therefore, there are no incentives for strategic types to behave honestly.

The link between individual and collective reputation is also important for explaining the persistence of corruption. A short-run increase in corruption, due, for example, to a bubble in the financial or housing markets, may destroy collective reputation, and have long-lasting effects. By the same token, effective anti-corruption policies have to be enforced for a long time in order to allow the group to rebuild a reputation for honesty. This difference may help to explain why corruption seems a more stable equilibrium.

8 EMPIRICAL EVIDENCE ON CORRUPTION

This section covers the most relevant academic empirical evidence related to corruption. In particular we analyze the measurement of corruption, the empirical evidence on the alternative models discussed in the previous sections and the evidence on anti-corruption mechanisms. We distinguish three basic sources of empirical results: the cross-country macro-evidence; the results derived from field experiments; and the evidence derived from lab experiments.

8.1 Measurement Issues

The question of measurement of corruption is complex since illegal activities, by definition, are secret transactions and, therefore, difficult to observe. The advantage of laboratory experiments on corruption is precisely the possibility to observe corruption directly in a controlled environment and at the level of an individual decision maker. Outside the lab corruption can be measured by at least three alternative methods: using perceptions, trying direct measures or using indirect methods.

8.1.1 Measuring the perception of corruption

Using the perception of corruption avoids the basic problem of many direct measurement methods that imply that bureaucrats/firms should disclose their participation in corrupt activities. Obviously, people are afraid to talk about their participation in illegal activities due to fear or shame. The admission of the perception of a high degree of corruption does not bear any negative connotation/feeling on the informants. Most of the first-generation studies on the determinants of corruption rely on cross-country data on the perception of corruption using macro-variables as explanatory variables. One of the first indicators of perception of corruption was constructed by the private company Business International and then incorporated by the Economist Intelligence Unit as part of the International Business Indicators Mauro (1995, p. 684), in the most cited article of the first generation of papers on the determinants of institutional efficiency, analyzes the determinants of corruption as defined by Business International: “the degree to which business transactions involve corruption or questionable payments.” Mauro (1995) finds that corruption reduces investment and, therefore, economic growth.²³ More recently Kaufmann, Kraay and Mastruzzi (2010)

²³ Although when corruption is instrumented with ethnic fractionalization the relationship is much weaker. Other well-known first-generation studies on the macro-determinants of corruption include Knack and Keefer (1995) and La Porta et al. (1999). Serra (2006) analyzes the robustness of the empirical findings based on subjective measures of corruption.

developed other measures of subjective corruption in a project sponsored by the World Bank. The Worldwide Governance Indicators (WGIs) of the World Bank cover 215 countries for 1996 to 2014.²⁴ The World Bank compiles six dimensions of governance including the control of corruption. The latest estimation of corruption relies on an average of nine different data sources for each of the 209 countries included in the study.²⁵ Another popular macro-indicator is the Global Corruption Barometer of Transparency International.

A recurrent finding of the empirical evidence of first-generation indicators, which is also common to studies based on non-subjective measures of corruption, is the negative relationship between corruption and proxies of economic development. However, the direction of causality is still a matter of debate. In addition, these subjective indicators usually present the point of view of experts who do not represent the perception of the whole population.²⁶ Finally, the most important problem of these measures is the fact that they correspond to perception of corruption and not actual corruption. Olken (2009) is able to estimate quite accurately the amount of actual corruption in a road project. He also gathered the assessment of the people living in the area with respect to the probability of corruption in the project. The correlation between both measures was very low, which implies that actual corruption was not well approximated to by the perception of corruption. In addition, Olken (2009) finds correlation between the perception of corruption and demographic characteristics, which is especially pervasive for the measurement based on perception when we have non-random samples such as the ones based on analysts or experts.

8.1.2 Direct estimates

An alternative to perceived corruption is to try to measure directly corruption activities. The direct estimation implies methods that could involve direct observation of bribes, audits or the use of surveys based on hypothetical situations. Olken and Barron (2009) measure directly the bribes of truck drivers to public officers on the roads of the province of Aceh in Indonesia. They use a simple method: surveyors dressed as assistants to truck drivers, take note of over 6,000 illegal payments to soldiers, police, etc. on 304 trips. Olken and Barron (2009) find that bribes are a very significant part of the cost of the trip: illegal payments amounted to 13 percent of the marginal cost of the trip compared with the 10 percent of the salary of the truck driver. Sequeira and Djankov (2014) analyze the illegal payments to avoid clearing fees (ex. tariff duties) in two ports: Maputo and Durban. They construct a dataset from direct observation of bribe payments to port officials for a random sample of 1,300 imports and 120 companies from South Africa. Sequeira and Djankov (2014) find that bureaucrats force private companies to pay fees above the official price of clearing services in 53 percent of the shipments to Maputo and the 34 percent of all shipments to Durban. However, the mean bribe was much higher in Maputo (14 percent of the shipment cost of a standard container) than in Durban (4 percent).²⁷ Bertrand et al. (2007) use data obtained by following 822 driver's license candidates in India and collected information on whether the license was also obtained at the time, the specific

²⁴ The latest version available at the time of writing in January 2016.

²⁵ The only relevant country for which there is no estimation is Monaco. For a methodological review see Kaufmann et al. (2010).

²⁶ In many countries it is possible to find subjective indicators of corruption, based, for instance, on the importance of corruption as a social problem, for representative samples of the population.

²⁷ McMillan and Zoido (2004) analyze the detailed records of the illegal activities of Montesinos, intelligence chief of President Fujimori of Peru. The size of the bribes ranged between 3,000 dollars per month to politicians to 1.5 million to TV stations.

processes and the expenditure. The experimental design compares a “bonus group,” which was offered a large financial reward if they could get the license in 32 days, and the “lesson group” who were offered free driving lessons. There was also a comparison group with no particular treatment, who were followed during the process. In this comparison group, close to 71 percent of those who obtained a license did not take the exam and 62 percent were not qualified to drive (tested by an independent examiner) at the time they got the license. The individuals in the comparison group also paid about 2.5 times the official fee to obtain a license. Individuals in the bonus group were 13 percentage points more likely to get the license without taking the driving exam and 18 percentage points more likely to obtain the license and, at the same time, fail the independent test.

The use of audits from watchdog institutions is also a methodology that could help to estimate corruption. Ferraz and Finan (2008, 2011) use official audits of municipalities in Brazil to construct a measure of political corruption in local governments. They find that around 8 percent of the total amount audited was diverted to illegal activities.²⁸ Duflo, Hanna and Ryan (2012) also use administrative data to measure teacher absenteeism in rural India.

In principle it is also possible to use surveys to measure bribes through questionnaires. For this purpose you need to find a procedure to encourage truth telling and carefully word your questions. The clearest example is Svensson’s (2003) study on the bribes paid by companies in Uganda. The basic problem with this methodology is that, no matter how careful you are in the preparation of the survey and the questionnaire, there is a high probability of underreporting.

8.1.3 Indirect estimates

These methods try to estimate illegal activities by comparing a magnitude using two or more sources that should lead to the same amount in absence of bribes.²⁹ Perhaps the best-known examples of these techniques are the public expenditure tracking surveys (PETS).³⁰ The PETS are quantitative surveys that collect information on facilities’ characteristics (mostly schools or clinics), financial flows and outputs. They trace the flow of resources from origin to destination and, therefore, PETS are well suited to uncovering any difference between the amount of resources provided by, for instance, a governmental agency, and the resources that reach the final destination (frontline facility). Reinikka and Svensson (2004) use a PETS to analyze the leakage rate in the block grant sent by the central government to schools in Uganda. They estimate that the leakage rate is a surprising 87 percent. Despite their potential to measure public fund leakages the low book-keeping quality of frontline facilities in developing countries may generate an upward bias in the estimations of leakages from PETS. Olken (2007) also uses a cross-checking technique to estimate the “missing expenditure” in the cost of rural roads projects. He compares the official project costs with the estimated cost by independent engineers, considering the same specifications. Olken (2007) concludes that 24 percent of the cost of the roads was “missing expenditure.” Fisman and Wei (2004) calculate “missing imports” as the difference between Hong Kong reported exports to China and the reported imports by China. They find that the rate of evasion was highly correlated with the tax rates of the products. In highly taxed products the evasion rate could reach 40 percent.

²⁸ Other research using administrative data includes Banerjee, Hanna and Mullainathan’s (2008) study of health centers’ charges and Atanassova, Bertrand and Mullainathan’s (2008) analysis of prices paid for goods that officially should be free. Both studies refer to India.

²⁹ For this reason some authors refer to this method as cross-checking.

³⁰ Montalvo (2003) describes and compares the methodologies used for different PETS.

Other cross-checking techniques are based on the comparison of administrative data and household surveys³¹ or the comparison between official prices and market prices. The best example of this second technique is Hsieh and Moretti's (2006) study of the UN Oil-for-Food program (1997–2003) to export Iraqi oil in exchange for food, medicine etc. for ordinary Iraqi citizens. They argue that Iraq set prices of oil below market prices to ask for bribes from buyers. Their estimation of corruption in this program amounts to US\$1.3 billion or 2 percent of oil revenues. Another well-known example is Fisman (2001) who estimates the value of political connections to President Soeharto of Indonesia as the difference between accounting data of the companies and stock prices. Fisman (2001) finds that 23 percent of the market value of the connected companies is the result of political ties with the dictator. Obviously this type of estimation relies heavily on the assumption of efficient markets.

8.1.4 Experimental estimates

Another way of dealing with the measurement of corruption and finding mechanisms to control it is to use lab experiments. Experimental data can avoid some of the problems found in the interpretation of other measures of corruption based on field experiments. In particular, you can avoid many endogeneity concerns (omitted variable bias, bidirectional causality, etc). Experiments also allow the analysis of the determinants of corruption among individuals and not only statements about aggregated behavior. Obviously, lab experiments are not a panacea. On many occasions it is difficult to justify the external validity of experiments.³²

There are many ways in which one can simulate a corrupt environment using a lab experiment. A frequent approach is to perform a game in which the higher payoff of the briber and the bribee produce a negative externality in the other players of the game that is larger than the sum of the private gains of briber and bribee. A good example of this type of experiment is presented in Abbink, Irlenbusch and Renner (2002). Depending on the type of corrupt activity one wants to simulate the game may be one-shot or repeated (for instance, corruption in procurement) or may involve only subjects playing the game or also passive subjects that perform activities not directly related to the experiment. This second alternative avoids the convoluted issue of forming beliefs about players' choices.

9 FIGHTING AGAINST CORRUPTION: THE EMPIRICAL EVIDENCE

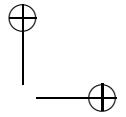
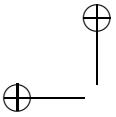
In this section we discuss the findings, derived from the empirical evidence, on the determinants of corruption and possible mechanisms to combat it. We consider two aspects already discussed from a theoretical perspective in the previous pages: the individual incentives of bureaucrats and the industrial organization aspects of competition in the briberies' market.

9.1 Bureaucratic Incentives: Punishment, Monitoring, Compensation, and Selection

We pointed out before that the most general view on the issue of regulation and corruption implies the analysis of a trade-off between the inefficiencies associated with regulation and the

³¹ Olken (2006) finds that at least 18 percent of the rice in a subsidized rice program in Indonesia can be qualified as "missing rice."

³² Armantier and Boly (2012) discuss this issue by comparing the results of a lab and a field experiment.



costs of the market failure that a particular regulation wants to correct. Bribes are a potential cost of public intervention. Gorodnichenko and Peter (2007) show that Ukrainian public workers had levels of consumption similar to private sector workers although their salary was much lower. Corruption seems to give extra income to public workers and governments can reduce their wages to offset the bribes they receive. In this case corruption could potentially lead to efficiency gains if the deadweight loss generated by the increase in taxation needed to increase salaries of public workers is larger than the deadweight loss coming from corruption. Although this trade-off is, in principle, theoretically relevant, it is, with very few exceptions, empirically less interesting. Therefore, we are going to assume that corruption is always an activity that is worth combating and that generates a net deadweight loss. Following Becker and Stigler's (1974) model, the basic condition for corruption in equilibrium is

$$w - \bar{w} < \frac{1 - \theta}{\theta} b,$$

where w is the wage of the bureaucrat, \bar{w} is the opportunity wage,³³ θ is the audit intensity and b is the size of the bribe. This last component, the size of the bribe, could be weighed against the psychological discomfort of dishonesty or b could be interpreted as the net effect. Following this simple theory to reduce corruption we can increase compensation of the bureaucrats, increase the penalties if identified (reduce the outside option), increase the probability of detection, or improve the selection criteria (increase the cost of dishonesty).

9.1.1 Delegation, monitoring and punishment

In principle increasing monitoring, and therefore the probability of detection, or increasing punishment should reduce corruption. Obviously, it could also be the case that more monitoring implies simply a redistribution of bribes between officials at different levels of the government (low ranking government workers, monitoring agents and workers at the sanctioning bodies).

Olken (2007) presents a convincing randomized field experiment on the effect of monitoring on the level of corruption using Indonesian road projects. He finds that increasing the probability of auditing from 4 percent to 100 percent reduces the discrepancies between official project cost and independently assessed cost of the projects by 8 percentage points. By contrast, grassroots monitoring in the form of citizens monitoring officials,³⁴ had little effect. It reduced the theft of wages but did not alter the theft of materials.

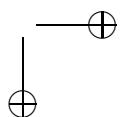
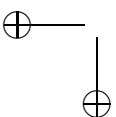
Ferraz and Finan (2011)³⁵ show that political institutions can affect corruption by increasing political accountability. They find significantly less corruption in municipalities where majors can be re-elected. In those cases the re-election incentive reduces 27 percent the misappropriated resources by 27 percent. There is also a large literature on lab experiments testing the effect of monitoring and punishment on corruption.³⁶ Abbink et al. (2002) find that when the probability of detection is low but the sanctions are very high, that being caught once implies very harsh consequences in terms of sanctions, their experiment shows a strong deterrence effect (reduces bribe offers by one-third). They conclude that the old fashion

³³ The difference between w and \bar{w} is also referred to as "efficiency wage."

³⁴ Olken (2007) examines two alternative interventions.

³⁵ Building on Ferraz and Finan (2008).

³⁶ Abbink and Serra (2012) present a lengthy discussion of this issue.



top-bottom vision of detection and harsh punishments as an effective way to combat corruption is correct. Schulze and Frank (2003), however, using a very different experimental set-up, find that in the no-risk case, a fixed amount independent of the bribe does not have any effect on corruptibility. In contrast, in the treatment with risk a fixed payment reduces the propensity of accept bribes. Abbink and Serra (2012) discuss several reasons for these contradictory findings although more research is needed to settle this issue.³⁷

A different institutional structure is monitoring bottom-up. For this purpose the system needs to impose some level of transparency that allows citizens and interested parties to hold public officials accountable for their decisions. One basic element for this system to work is transparency. Improving public information is critical to any bottom-up strategy. If citizens can monitor public employees then electoral accountability could potentially prevent corruption.³⁸ In addition, providing information also empowers citizens to complain when the public programs intended to improve their living conditions are not appropriately developed. Reinikka and Svensson (2005) use a PETS study in Uganda as the set-up for a policy experiment in which the government ran a newspaper campaign for parents. The original PETS study (Reinikka and Svensson, 2004) found that schools received only 20 percent of the funds transfer to local authorities by the central government. To reduce the diversion of public money the government of Uganda provided information to parents and schools that could help them to monitor the local officials' handling of a large school grant program. Reinikka and Svensson (2005) use the distance to the nearest newspaper selling point as an instrumental variable. They found that there was a strong relationship between the proximity of the kiosk and the reduction of the diversion of school funds. They also report an increase in enrolment and learning.

Another popular transparency strategy to prevent corruption is to publicize the initial level of wealth of politicians to provide a yardstick of comparison with the wealth when they leave office. There is little research on this issue. One of the few examples is Djankov et al. (2010) who collected cross-country information on the rule of financial and conflict disclosure of Members of Parliament. They find that public disclosure is associated with less corruption.

In the previous paragraphs we have discussed monitoring using first a top-down approach and last a bottom-up strategy. Serra (2011) proposes a lab experiment that combines both approaches. She concludes that the "combined" system of accountability is highly efficient against corruption.

9.1.2 Bureaucrats and compensation

Should compensation of public servants be high to prevent the temptation of corruption? This is a much debated issue. Using cross-country data Rauch and Evans (2000) show that bureaucrats' wages are only significant in the explanation of one out of five measures of bureaucratic performance (the Bureaucratic Delay Index). However, Van Rijckeghem and Weder (2001) find a significant negative effect of public wages on the International Country Risk Guide's Index of Corruption and, therefore, conclude that higher salaries lead to lower corruption. Le et al. (2013) argue that the effect of public wages on the prevention of corruption depends on the level of development of the countries. Using a new micro-level

³⁷ Abbink and Serra (2012) also discuss monitoring through the four eyes principle, which implies that public service decisions should be made by more than one official.

³⁸ Ferraz and Finan (2008) find that the dissemination of information on audits that show corruption in local governments had negative electoral effect on incumbents in local elections in Brazil.

dataset and controlling for a number of other determinants of corruption and country fixed effects they show that the effectiveness of public wages to reduce corruption decreases with the level of development. However, Besley and McLaren (1993) and Macchiavello (2008) argue that higher public wages are not a good device to reduce corruption in less developed countries since it can generate a bad selection effect: highly motivated workers can be crowded out by individuals susceptible to corruption.

From a microeconomic perspective Di Tella and Schargrodsy (2003) find an interaction between monitoring intensity and public wages in their analysis of the corruption activities related to procurement in hospitals of Buenos Aires. They conclude that the degree of audit intensity is fundamental for the effectiveness of wage policies against corruption. High public wages are only effective if there is sufficient auditing intensity. By contrast, the effective of intensive auditing may not be sustainable over time.

The experimental literature has also debated the issue of public earnings and its deterrence effect on corruption. Using the set-up of Abbink et al. (2002) but assuming that the negative externality affects individuals doing tasks unrelated to the experiment, Abbink (2004) finds no effect of higher wages. Van Veldhuizen (2013) modifies the set-up of Abbink et al. (2002) to inflict the negative externality of corruption on a donation that the experimenter should make to a charity. He finds that increasing public officials' wages reduces corruptibility: low paid officials accept bribes 91 percent of the time while among highly paid officials the proportion is reduced to 38 percent. Van Veldhuizen (2013), similar to the findings of Di Tella and Schargrodsy (2003), shows that a positive monitoring rate is necessary to find this negative effect of high wages on corruption.

Armantier and Boly (2012) use the comparison of a field and a lab experiment and find that in both set-ups higher paid graders of students' exams had a lower probability of accepting bribes.

9.1.3 Selection

The determinants of the recruitment of public sector workers are a less developed topic of research. Dal Bó, Finan and Rossi (2013) use a randomized field experiment to analyze the role of financial incentives in the recruitment of government officials. They use the Regional Development Program of Mexico. The objective was to enhance the presence of the Mexican state in 167 marginal municipalities by hiring 350 community development agents who had to identify deficiencies in the provision of public goods. The hiring of these agents was conducted using an exogenous allocation of wage offers and job offers across recruitment sites. Dal Bó et al. (2013) find that higher wages attracted more able applicants as measured by their IQ and proclivity towards government work. In addition, higher wage offers increase acceptance rates. Ferraz and Finan (2010) show that higher salaries increased political competition and attract more educated candidates, although the effect was small in general. Higher wages also improved the performance of politicians.

9.1.4 Incentives and institutional structures for monitoring

Incentives can reduce corruption by linking performance to pay. Much of the research on this issue has been directed towards the provision of education and health in less developed countries. A basic problem for the frontline provision of education and health is absenteeism. In many of those countries teachers and doctors are also a powerful civil force able to avoid the pressure to attend their obligations. Duflo et al. (2012) use a randomized experiment to

test the effect of monitoring and financial incentives on teacher absenteeism in rural India. The treatment consisted of monitoring using cameras and salaries were a function of attendance. The absenteeism of the treated group fell by 21 percentage points relative to the control group. Increasing teacher attendance does not necessarily increase child learning since it can affect intrinsic motivation to teach and demoralize teachers. Duflo et al. (2012) also found that children's test scores had increased by 0.17 standard deviation points in the treatment group.³⁹ Banerjee, Duflo and Glennerster (2008) report on a randomized evaluation of an incentive program to increase the attendance at rural health centers in India. The evaluation uses timeclocks to monitor nurses' attendance, which, in turn, determined their wages. The system was initially very efficient and showed that nurses reacted to financial incentives. However, political economy considerations (nurses' pressure) undermined the program that, after one-and-a-half years, became ineffective.

9.2 Competition and the market for bribes

We have already discussed in the theory section the “double-marginalization” problem that can appear if there are multiple corrupt officials (Shleifer and Vishny, 1993). Olken and Barron (2009) use a natural experiment to test this theory, exploiting the reduction in the checkpoints at which trucks in Aceh had to stop to pay bribes due to a peace agreement with a rebel group. They show an increase in the average bribes observed on the check-points on the remainder as a reaction to the withdrawal of checkpoints in Aceh. In addition, Olken and Barron (2009) also show that the bribes increase the closer the truck is to the destination, which supports the idea of an ex post hold-up in a chain of monopolies.

Burgess et al. (2012) study the effect of competition between bureaucrats in the context of deforestation in Indonesia. From 1998 to 2008 the number of districts increased from 292 to 483, while during the 32 years of the Soeharto regime they remained mostly unchanged. This increase in the level of decentralization of the government also affected the management of the forests. In principle, there were logging quotas but companies could go beyond the legal limit by paying bribes to the district officials. Burgess et al. (2012) find that the increase in districts lead to increased deforestation and lower timber prices consistent with Cournot competition between districts' officials. They also find short-run substitution between alternative forms of corruption, in particular illegal logging and rents from oil and gas, although this effect disappears over time. This substitution effect in the short run mimics that found by Olken (2007) after the introduction of audits: the diverted expenditures decreased but the nepotism increased.

Drugov, Hamman and Serra (2014) study the effect of intermediaries as facilitators of corruption activities. Although their experimental set-up cannot be easily compared with a particular market structure it shows the result of introducing multiple layers in a bribery experiment. Drugov et al. (2014) conduct a one-shot petty bribery game in which they change the number of intermediaries. They find that the proportion of corrupt briber-bribee couples increases significantly when there are intermediaries and argue that intermediaries reduce moral or psychological cost and, therefore, increase corruption.

In the theoretical part of the chapter we also discussed how job rotation can have a deterrence effect on corruption by making reputation building more difficult and revealing

³⁹ Muralidharan and Sundararaman (2011) also find positive effects of financial incentives on students' average test scores.

corrupt types. Staff rotation has been a standard procedure in some industries like the financial sector. Branch directors are regularly moved to a different location to avoid the development of personal relationships with clients that can affect the profitability of the bank. There is little empirical evidence on the effect of staff rotation. Abbink (2004), using the set-up of Abbink et al. (2002), includes a random reshuffling of pairs of potential bribers and public officials in every round. The rotation reduces the frequency of inefficient decisions due to bribery by two-thirds and bribes, on average, are reduced by almost one-half. Needless to say that the high effectiveness of staff rotation could be the result of the particular experiment used to replicate the environment of a corruptible relationship and, in any case, it is not a cost–benefit analysis. It does not consider the cost of moving a worker with experience in a particular job/location to a different job/location.

In the theory section we also show that generating incentives for parties to report wrongdoing is a potential strategy to reduce the cost of monitoring corruption. However, as we argued, leniency policies whereby a reporting party may be exempted from penalties, may make infeasible briberies possible. To avoid this effect it is possible to design an asymmetric system of penalties and leniency program. Schikora (2011) finds, using a lab experiment, that symmetrically punished whistle-blowing has an ambiguous effect as predicted by the literature because although it reduces the impact of corruption it increases its stability. However, he finds that asymmetric leniency reduces corruption because government officials have the opportunity to avoid the threat of retaliation by the client without the risk of penalization by whistle-blowing.

10 CONCLUSIONS

In this chapter we present a lengthy discussion on theoretical and empirical issues related to corruption. In the final sections we discuss empirical issues. First of all we analyze the question of the measurement of corruption. This is a complex task since illegal activities, by definition, are secret transactions and, therefore, difficult to observe. We cover all the available methodologies: direct and indirect estimation, perceived corruption and experimental elicitation of the propensity to engage in illegal activities. Using these alternative measures of corruption we then turn to the analysis of the predictive ability of the theoretical models given the available empirical evidence. Many of the mechanisms described in the theoretical models have been found to be relevant in empirical applications. There is strong evidence on the effect of monitoring and punishment on the extension of corruption. There is also increasing evidence on the “double-marginalization” effect caused by the presence of multiple corrupt officials. There is less evidence on the effect of compensation on the behavior of bureaucrats. More empirical research is needed on the specific mechanisms that can be effective to deter corruption and illegal activities.

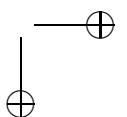
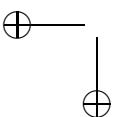
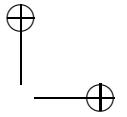
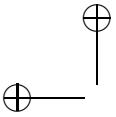
REFERENCES

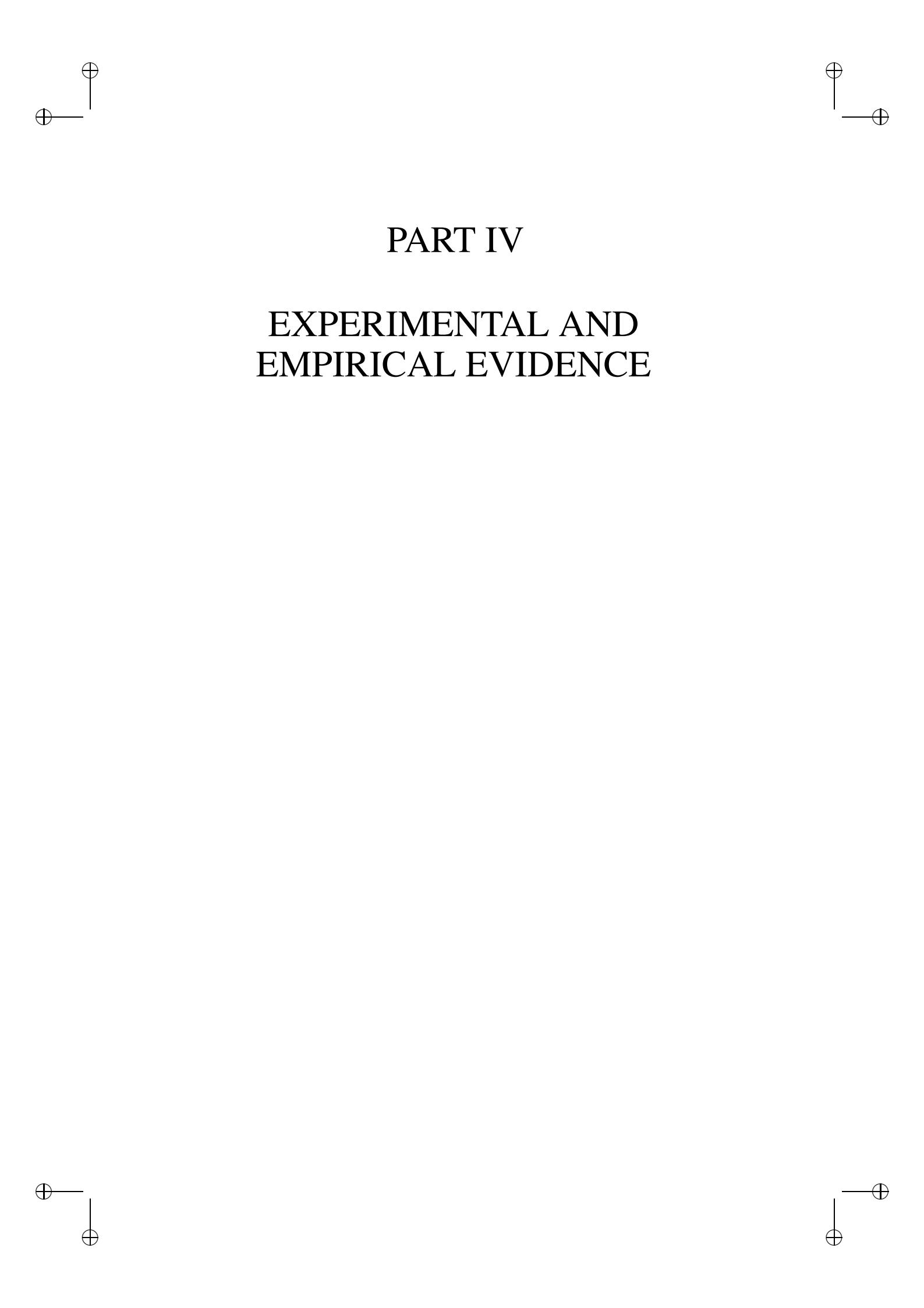
- Abbink, K. (2004). “Staff rotation as an anti-corruption policy: an experimental study,” *European Journal of Political Economy*, Vol. 20, pp. 887–906.
 Abbink, K. and D. Serra (2012). “Anticorruption policies: lessons from the lab,” in D. Serra and L. Wantchekon (eds), *New Advances in Experimental Research on Corruption, Research in Experimental Economics. Volume 15*, Bingley, UK: Emerald Group Publishing.

- Abbink, K., Irlenbusch B., and E. Renner (2002), "An experimental bribery game," *Journal of Law, Economics, and Organization*, Vol. 18, No. 2, pp. 428–454.
- Acemoglu, D. and T. Verdier (2000), "The choice between market failures and corruption," *American Economic Review*, Vol. 90, No. 1, pp. 194–211.
- Ades, A. and R. Di Tella (1999), "Rents, competition, and corruption," *American Economic Review*, Vol. 89, No. 4, pp. 982–993.
- Akerlof, G.A. (1970), "The market for 'Lemons': quality uncertainty and the market mechanism," *Quarterly Journal of Economics*, Vol. 84, pp. 488–500.
- Amir, R. and C. Burr (2015), "Corruption and socially optimal entry," *Journal of Public Economics*, Vol. 123, pp. 30–41.
- Andvig, J. and K. Moene (1990), "How corruption may corrupt," *Journal of Economic Behavior and Organization*, Vol. 13, No. 1, pp. 63–76.
- Armantier, O. and A. Boly (2012), "On the external validity of corruption experiments," in D. Serra and L. Wantchekon (eds), *New Advances in Experimental Research on Corruption, Research in Experimental Economics*. Volume 15, Bingley, UK: Emerald Group Publishing.
- Arozamena, L. and F. Weinschelbaum (2009), "The effect of corruption on bidding behavior in first-price auctions," *European Economic Review*, Vol. 53, No. 6, pp. 645–657.
- Atanassova, A., M. Bertrand, and S. Mullainathan (2008), "Misclassification in targeted programs: a study of the targeted public distribution system in Karnataka, India," mimeo.
- Auriol, E. (2006), "Corruption in procurement and public purchase," *International Journal of Industrial Organization*, Vol. 24, pp. 867–885.
- Auriol, E. and T. Soreide (2015), "An economic analysis of debarment," mimeo.
- Baliga, S. and T. Sjostrom (1998), "Decentralization and collusion," *Journal of Economic Theory*, Vol. 83, pp. 196–232.
- Banerjee, A., E. Duflo, and R. Glennerster (2008), "Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system," *Journal of the European Economic Association*, Vol. 6, pp. 487–500.
- Basu, K., K. Basu, and T. Cordella (2014), "Asymmetric punishment as an instrument of corruption control," *Policy Research Working Paper* 6933, The World Bank.
- Beck, P. and W. Maher (1986), "A comparison of bribery and bidding in thin markets," *Economics Letters*, Vol. 20, No. 1, pp. 1–5.
- Becker, G.S. and G.J. Stigler (1974), "Law enforcement, malfeasance, and compensation of enforcers," *Journal of Legal Studies*, Vol. 3, No. 1, pp. 1–18.
- Bertrand, M., S. Djankov, R. Hanna, and S. Mullainathan (2007), "Obtaining a driver's license in India: an experimental approach to studying corruption," *Quarterly Journal of Economics*, Vol. 122, pp. 1639–1676.
- Besley, T. and J. McLaren (1993), "Taxes and bribery: the role of wage incentives," *Economic Journal*, Vol. 103, No. 416, pp. 119–141.
- Bliss, C. and R. Di Tella (1997), "Does competition kill corruption?" *Journal of Political Economy*, Vol. 105, No. 5, pp. 1001–1023.
- Buccirossi, P. and G. Spagnolo (2006), "Leniency policies and illegal transactions," *Journal of Public Economics*, Vol. 90, pp. 1281–1297.
- Burgess, R., M. Hansen, and B. Olken et al. (2012), "The political economy of deforestation in the tropics," *Quarterly Journal of Economics*, Vol. 127, No. 4, pp. 1707–1754.
- Burguet, R. (2014), "Procurement design with corruption," *Barcelona GSE Working Paper Series* 798.
- Burguet, R. and Y.K. Che (2004), "Competitive procurement with corruption," *The RAND Journal of Economics*, Vol. 35, No. 1, pp. 50–68.
- Burguet, R. and M.K. Perry (2007), "Bribery and favoritism by auctioneers in sealed-bid auctions," *The B.E. Journal of Theoretical Economics*, Vol. 7, No. 1.
- Cadot, O. (1987), "Corruption as a gamble," *Journal of Public Economics*, Vol. 33, No. 2, pp. 223–244.
- Cameron, L., A. Chaudhuri, N. Erkal, and L. Gangadharan (2009), "Propensities to engage in and punish corrupt behavior: experimental evidence from Australia, India, Indonesia and Singapore," *Journal of Public Economics*, Vol. 93, No. 7–8, pp. 843–851.
- Celentani, M. and J.J. Ganuza (2002), "Corruption and competition in procurement," *European Economic Review*, Vol. 46, No. 7, pp. 1273–1303.
- Celik, G. (2009), "Mechanism design with collusive supervision," *Journal of Economic Theory*, Vol. 144, pp. 69–95.
- Chassang, S. and G. Padró-i-Miquel (2014), "Corruption, intimidation, and whistle-blowing: a theory of inference from unverifiable reports," *NBER Working Paper* No. 20315.
- Choi, J.P. and M. Thum (2003), "The dynamics of corruption with the ratchet effect," *Journal of Public Economics*, Vol. 87, pp. 427–443.
- Choi, J.P. and M. Thum (2005), "Corruption and the shadow economy," *International Economic Review*, Vol. 46, No. 3, pp. 817–836.

- Compte, O., A. Lambert-Mogiliansky, and T. Verdier (2005), "Corruption and competition in procurement auctions," *The RAND Journal of Economics*, Vol. 36, No. 1, pp. 1–15.
- Dal Bó, E., F. Finan, and M. Rossi (2013), "Strengthening state capabilities: the role of financial incentives in the call to public service," *Quarterly Journal of Economics*, Vol. 128, No. 3, pp. 1169–1218.
- Dechenaux, E. and A. Samuel (2012), "Pre-emptive corruption, hold-up and repeated interactions," *Economica*, Vol. 79, pp. 258–283.
- De Soto, H. (1989), *The Other Path*, New York: Harper & Row.
- Di Tella, R. and E. Schargrodsky (2003), "The role of wages and auditing during a crackdown on corruption in the city of Buenos Aires," *The Journal of Law and Economics*, Vol. 46, No. 1, pp. 269–292.
- Djankov, S., R. La Porta, F. Lopez de Silanes, and A. Shleifer (2010), "Disclosure by politicians," *American Economic Journal: Applied Economics*, Vol. 2, No. 2, pp. 179–209.
- Drugov, M. (2010), "Competition in bureaucracy and corruption," *Journal of Development Economics*, Vol. 92, No. 2, pp. 107–114.
- Drugov, M., J. Hamman, and D. Serra (2014), "Intermediaries in corruption: an experiment," *Experimental Economics*, Vol. 17, pp. 78–99.
- Duflo, E., R. Hanna, and S. Ryan (2012), "Incentives work: getting teachers to come to school," *American Economic Review*, Vol. 102, No. 4, pp. 1241–1278.
- Dufwenberg, M. and G. Spagnolo (2015), "Legalizing bribe giving," *Economic Inquiry*, Vol. 53, No. 2, pp. 836–853.
- Ferraz, C. and F. Finan (2008), "Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes," *Quarterly Journal of Economics*, Vol. 123, pp. 703–745.
- Ferraz, C. and F. Finan (2010), "Motivating politicians: the impacts of monetary incentives on quality and performance," unpublished manuscript, University California Berkeley.
- Ferraz, C. and F. Finan (2011), "Electoral accountability and corruption: evidence from the audit reports of local governments," *American Economic Review*, Vol. 101, pp. 1274–1311.
- Fisman, R. (2001), "Estimating the value of political connections," *American Economic Review*, Vol. 91, pp. 1095–1102.
- Fisman, R. and S.J. Wei (2004), "Tax rates and tax evasion: imports in China," *Journal of Political Economy*, Vol. 112, pp. 471–496.
- Gorodnichenko, Y. and K.S. Peter (2007), "Public sector pay and corruption: measuring bribery from micro data," *Journal of Public Economics*, Vol. 91, pp. 963–991.
- Hsieh, C.R. and E. Moretti (2006), "Did Iraq cheat the United Nations? Underpricing, bribes, and the Oil for Food Program," *Quarterly Journal of Economics*, Vol. 121, pp. 1211–1248.
- Kaufmann, D., Kraay, A., and M. Mastruzzi (2010), "The Worldwide Governance Indicators: methodology and analytical issues," *Policy Research Working Paper 5430*.
- Kessler, A.S. (2000), "On monitoring and collusion in hierarchies," *Journal of Economic Theory*, Vol. 91, No. 2, pp. 280–291.
- Khalil, F., J. Lawarée, and S. Yun (2010), "Bribery versus extortion: allowing the lesser of two evils," *RAND Journal of Economics*, Vol. 41, No. 1, pp. 179–198.
- Klitgaard, R. (1988), *Controlling Corruption*, Berkeley, CA: University of California Press.
- Knack, S. and P. Keefer (1995), "Institutions and economic performance: cross-country tests using alternative institutional measures," *Economics & Politics*, Vol. 7, No. 3, pp. 207–227.
- Koc, S.A. and W.S. Neilson (2008), "Interim bribery in auctions," *Economics Letters*, Vol. 99, No. 2, pp. 238–241.
- Kofman, F. and J. Lawarée (1996), "On the optimality of allowing collusion," *Journal of Public Economics*, Vol. 61, pp. 383–407.
- Laffont, J.J. and T. N'Guessan (1999), "Competition and corruption in an agency relationship," *Journal of Development Economics*, Vol. 60, pp. 271–295.
- Laffont, J.J. and J. Tirole (1991), "The politics of government decision-making: a theory of regulatory capture," *Quarterly Journal of Economics*, Vol. 196, No. 4, pp. 1089–1127.
- Lambert-Mogiliansky, A., M. Majumdar, and R. Radner (2007), "Strategic analysis of petty corruption: entrepreneurs and bureaucrats," *Journal of Development Economics*, Vol. 83, No. 2, pp. 351–367.
- Lambert-Mogiliansky, A., M. Majumdar, and R. Radner (2008), "Petty corruption: a game-theoretic approach," *International Journal of Economic Theory*, Vol. 4, pp. 273–297.
- La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny (1999), "The quality of government," *The Journal of Law, Economics, and Organization*, Vol. 15, No. 1, pp. 222–279.
- Le, V., J. de Haan, and E. Dietzenbacher (2013), "Do higher government wages reduce corruption? Evidence from a novel dataset," *CESifo Working Paper 4254*.
- Lengwiler, Y. and E. Wolfstetter (2010), "Auctions and corruption: an analysis of bid rigging by a corrupt auctioneer," *Journal of Economic Dynamics and Control*, Vol. 34, No. 10, pp. 1872–1892.
- Lien, D. (1986), "A note on competitive bribery games," *Economics Letters*, Vol. 22, No. 4, pp. 2337–2341.
- Lui, F.T. (1986), "An equilibrium queuing model of bribery," *Journal of Political Economy*, Vol. 93, No. 4, pp. 760–781.

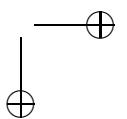
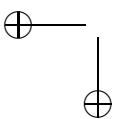
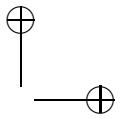
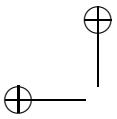
- Macchiavello, R. (2008), "Public sector motivation and development failures," *Journal of Development Economics*, Vol. 86, No. 1, pp. 201–213.
- Mauro, P. (1995), "Corruption and growth," *Quarterly Journal of Economics*, Vol. 110, pp. 681–712.
- McMillan, J. and P. Zoido (2004), "How to subvert democracy: Montesinos in Peru," *Journal of Economic Perspectives*, Vol. 18, No. 4, pp. 69–92.
- Menezes, F.M., and P.K. Monteiro (2006), "Corruption and auctions," *Journal of Mathematical Economics*, Vol. 42, No. 1, pp. 97–108.
- Montalvo, J.G. (2003), "Designing a PETS in education for the Republic of Philippines: lessons from previous PETS experiences," World Bank, mimeo.
- Monteiro, J. and C. Ferraz (2010), "Does oil make leaders unaccountable? Evidence from Brazil's offshore oil boom," unpublished manuscript, PUC-Rio.
- Mookherjee, D. and I.P.L. Png (1995), "Corruptible law enforcers: how should they be compensated?" *Economic Journal*, Vol. 105, No. 428, pp. 145–159.
- Muralidharan, K. and V. Sundararaman (2011), "Teacher performance pay: experimental evidence from India," *Journal of Political Economy*, Vol. 119, pp. 39–77.
- Olken, B.A. (2006), "Corruption and the costs of redistribution: micro evidence from Indonesia," *Journal of Public Economics*, Vol. 90, pp. 853–870.
- Olken, B.A. (2007), "Monitoring corruption: evidence from a field experiment in Indonesia," *Journal of Political Economy*, Vol. 115, pp. 200–249.
- Olken, B.A. (2009), "Corruption perceptions vs. corruption reality," *Journal of Public Economics*, Vol. 93, pp. 950–964.
- Olken, B.A. and P. Barron (2009), "The simple economics of extortion: evidence from trucking in Aceh," *Journal of Political Economy*, Vol. 117, pp. 417–452.
- Pechlivanos, L. (2005) "Self-enforcing corruption: information transmission and organizational response," in J. Graf Lambsdorff, M. Taube, and M. Schramm (eds), *The New Institutional Economics of Corruption*, London: Routledge.
- Rauch, J.E. and P.B. Evans (2000), "Bureaucratic structure and bureaucratic performance in less developed countries," *Journal of Public Economics*, Vol. 75, pp. 49–71.
- Reinikka, R. and J. Svensson (2004), "Local capture: evidence from a central government transfer program in Uganda," *Quarterly Journal of Economics*, Vol. 119, pp. 679–706.
- Reinikka, R. and J. Svensson (2005), "Fighting corruption to improve schooling: evidence from a newspaper campaign in Uganda," *Journal of the European Economic Association*, Vol. 3, pp. 259–267.
- Rose-Ackerman, S. (1978), *Corruption: A Study of Political Economy*, New York: Academic Press.
- Rose-Ackerman, S. (1996), "Redesigning the state to fight corruption," *Public Policy for Private Sector*, World Bank Note No. 75.
- Schickora, J.T. (2011), "Bringing good and bad whistle-blowers to the lab," *Discussion Papers in Economics*.
- Schulze, G.G. and B. Frank (2003), "Deterrence versus intrinsic motivation: experimental evidence on the determinants of corruptibility," *Economics of Governance*, Vol. 4, No. 2, pp. 143–160.
- Sequeira, S. and S. Djankov (2010) "An empirical study of corruption in ports," *Working Paper*, London School of Economics.
- Serra, D. (2006), "Empirical determinants of corruption: a sensitivity analysis," *Public Choice*, Vol. 126, No. 1–2, pp. 225–256.
- Serra, D. (2011) "Combining top-down and bottom-up accountability: evidence from a bribery," *Working Paper*.
- Shleifer, R. and R.W. Vishny (1993), "Corruption," *Quarterly Journal of Economics*, Vol. 108, No. 3, pp. 599–617.
- Shleifer, A. and R.W. Vishny (1994), "Politicians and firms," *Quarterly Journal of Economics*, Vol. 109, No. 4, pp. 995–1025.
- Strausz, R. (1997), "Collusion and renegotiation in a principal–supervisor–agent relationship," *The Scandinavian Journal of Economics*, Vol. 99, No. 4, pp. 497–518.
- Svensson, J. (2003), "Who must pay bribes and how much? Evidence from a cross-section of firms," *Quarterly Journal of Economics*, Vol. 118, pp. 207–230.
- Tirole, J. (1986), "Hierarchies and bureaucracies: on the role of collusion in organizations," *Journal of Law, Economics & Organization*, Vol. 2, No. 2, pp. 181–214.
- Transparency International (2006), *Handbook for Curbing Corruption in Public Procurement*, available at https://www.transparency.org/whatwedo/publication/handbook_for_curbig_corruption_in_public_procurement.
- Van Rijckeghem, C. and B. Weder (2001), "Bureaucratic corruption and the rate of temptation: do wages in the civil service affect corruption, and by how much?" *Journal of Development Economics*, Vol. 65, pp. 307–331.
- Van Veldhuizen, R. (2013), "The influence of wages on public officials' corruptibility: a laboratory investigation," *Journal of Economic Psychology*, Vol. 32, pp. 341–356.

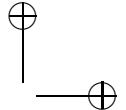
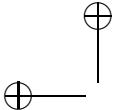




PART IV

EXPERIMENTAL AND EMPIRICAL EVIDENCE





17. Experimental industrial organization

Jordi Brandts and Jan Potters

1 INTRODUCTION

We present a selective survey of experimental studies on industrial organization (IO) issues centered on recent work. More material can be found in the seminal survey by Holt (1995), the meta-study on collusion by Engel (2007), the book on experiments and competition policy edited by Hinloopen and Normann (2009) and the survey on oligopoly experiments in the new millennium by Potters and Suetens (2013).

We believe that the results from laboratory experiments can help understand strategic behavior in general and in industrial organization settings in particular due to the twin virtues of experimentation: *control* and *replicability*.¹

Control has several dimensions. First, the experimenters can create the situation in the lab based on a particular game-theoretic model and, hence, know the relevant equilibrium (or equilibria) exactly. This includes aspects like information conditions (which players have which pieces of information) and exogenous stochastic processes, which in natural data are difficult to ascertain. Since modern industrial organization is inextricably linked to game-theoretic modeling, the fact that in the lab one can reproduce the conditions that define theoretical models makes lab experiments an important source of empirical knowledge in the area.

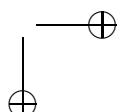
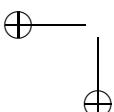
A second dimension of control involves the fact that the experimenters can make *ceteris paribus* changes in relevant variables. They can create counterfactuals. These changes are truly exogenous, since they are directly imposed by the experimenters. This facilitates causal inference and in many cases greatly simplifies the statistical analysis. Third, participants in experiments are randomly assigned to different treatment conditions, ruling out selection bias.

With respect to replicability we highlight two main dimensions. First, it is typically possible to generate sufficient data to be able to perform the appropriate statistical tests. Second, skeptics of particular experimental results can run their own experiments, when there are doubts about the exact conduct of the experiments or the truthfulness in reporting the results. Indeed, most debates in experimental economics are not settled by re-examination of existing data, but by conducting replication studies.

The chapter is organized as follows. Section 2 presents, starting with the classical models of Cournot, Bertrand and Stackelberg, results from experiments based on static models involving the choice of quantities and prices. Section 3 deals with tacit collusion. Section 4 covers (horizontal) product differentiation and Section 5 discusses experience and credence goods. Section 6 presents studies about entry deterrence and R&D competition. Section 7 concludes.²

¹ For a more detailed discussion of the usefulness of laboratory experiments in the social sciences see Falk and Heckman (2009) and for a comprehensive treatment of the methodology of experimental economics see Guala (2005).

² Space limitations force us to skip some areas of interest. These include experiments on consumer search (e.g., Davis and Holt, 1996; Cason and Friedman, 2003), price dispersion and advertising (Morgan, Orzen, and Sefton,



2 OLIGOPOLY MODELS: QUANTITIES AND PRICES

There is a strand of experiments based on static oligopoly models involving quantities and prices as decision variables. Central in this work are the experiments pertaining to the classic Cournot and Bertrand models of competition, as well as to the well-known Stackelberg model involving firms making decisions sequentially. However, there are also experiments in which quantities and/or prices are chosen in different ways than in the classic models.³ We first discuss the results pertaining to pure quantity and price competition. We then turn to the experimental evidence related to the Kreps-Scheinkman model where first quantities and then prices are chosen and subsequently to the setting where quantities are chosen simultaneously with prices, implying that production takes place before sales are determined. At the end of this section we discuss the evidence on supply function competition, where different units of output are offered at potentially different prices and a uniform market price is determined in a call auction.

We start with an important procedural issue. Typically the equilibrium models on which the experiments are based are static, i.e., they pertain to the one-shot play of the underlying game. However, in most experiments participants make choices over multiple rounds. This leads to an important distinction in experimental design. Some experiments involve repeated play with opponents that change from round to round, whereas in other experiments opponents are fixed over time. Having participants play a game repeatedly but with changing opponents is the procedure most commonly used by experimentalists to allow participants to familiarize themselves with the game and, at the same time, to remain close to the static nature of the equilibrium models behind the experiments.⁴ Studying repeated play with fixed partners is also of interest since in naturally occurring oligopoly markets it is typically a constant group of firms that interact over time. Researchers with a more applied inclination may be more interested in the results of interactions in fixed groups. In the case of repeated interaction an additional issue is whether there is a finite and known number of repetitions or the interaction takes place indefinitely. Indefinitely repeated interaction corresponds more closely to the way in which economic theory envisions repeated interaction. However, many experiments use finitely repeated interaction, according to an early less theory-inspired experimental tradition.

The extent to which play with changing opponents (a procedure often referred to as *strangers*) differs from that with fixed opponents (referred to as *partners*) is an issue overarching the different models of interaction. The well-known folk theorem (Aumann and Shapley, 1994) shows that in repeated games with uncertain duration, collusion (at different levels) is an equilibrium. For finitely repeated games there can be collusive equilibria under certain conditions about the stage game. From a theoretical point of view one-shot and repeated interaction are very different, but it is an empirical question as to what extent behavior will indeed be different in the two settings. In his meta-analysis of oligopoly

2006a, 2006b), and experiments on antitrust (mergers, predation, foreclosure, exclusive dealings). For excellent reviews on the latter literature see Goette and Schmutzler (2009), and Müller and Normann (2014).

³ Reinhard Selten's work marks the beginning of the experimental analysis of oligopoly. Indeed, his research includes both one of the first experiments on quantity competition and the first one on price competition. Sauermann and Selten (1959) and Hoggatt (1959) are the first experimental papers on quantity competition and Selten (1967) is the first paper on price competition. For a discussion of the early work on quantity and prices competition see Bosch-Domènec and Vriend (2008) and Abbink and Brandts (2010).

⁴ Playing repeatedly with changing partners is not the same as a one-shot interaction, since subjects learn and adapt over time, but it is seen as the most practical approximation to it.

experiments Engel (2007) directly compares the degree of collusion observed under strangers and under partners. The surprising result is that on average, strangers led to higher collusion than partners as measured by three distinct collusion indices. If one restricts the analysis to those papers that specifically compare strangers and partners, then there is no significant difference between strangers and partners. An interpretation of this absence of differences is that tacit collusion in repeated interaction is simply not easy to accomplish without a way of coordinating actions and that at the same time in one-shot interaction participants do somehow learn over time to strive for higher than equilibrium prices. Naturally this result of the meta-study does not mean that the distinction between one-shot and repeated interaction is unimportant, but it cautions against jumping to conclusions. Some of the experiments discussed in this section use a strangers protocol, whereas others use partners and again some others use both.

2.1 Quantity Competition and Price Competition

Holt (1985) contains some results from quantity competition duopolies with changing partners that show that average behavior is close to the Cournot equilibrium of the game. The situation is presented to participants as a simple matrix game. Dufwenberg and Gneezy (2000) study price competition with a homogeneous product and constant marginal cost with changing partners. They find that “the Bertrand solution does not predict well when the number of competitors is two, but (after some opportunities for learning) predicts well when the number of competitors is three or four” (p. 7). In summary, the equilibria of the basic static quantity and price competition games predict experimental behavior rather well, with the exception of the case of price competition with two competitors. This exception can be explained in terms of subjects’ resistance to ending up in a situation with zero profits; this resistance is somewhat successful with two competitors but fails with more than two.

For price competition the existence of firm capacity constraints changes the equilibrium of the game very substantially. Kruse et al. (1994) study price competition in four-player markets with exogenously given capacities and proportional demand rationing. For this case, there is no equilibrium in pure strategies. They find that prices are lower with higher capacities, but are higher than at the corresponding static Nash equilibrium. Their data also suggest that experimental participants adjust prices following a myopic best response, as suggested by Edgeworth. Other papers also find evidence of price cycles (Durham et al., 2004; Bruttel, 2009a; Peeters and Strobel, 2009; Leufkens and Peeters, 2010; Davis, 2011).

Davis and Holt (1994b) study how prices depend on the distribution of capacity among five sellers. In the baseline treatment total capacity is distributed such that the Nash equilibrium price is the Walrasian price. In a market power treatment total capacity is reallocated among the sellers in such a way that the mixed-strategy Nash equilibrium distribution has a mean above the Walrasian level. This theoretical market power is found to significantly raise prices in the experiments. Fonseca and Normann (2013) vary capacity levels and the number of market competitors. They find, consistent with Kruse et al. (1994), that higher capacities lead to lower prices and that price movements over time are more in line with the notion of Edgeworth cycles than with the static Nash equilibrium.⁵

⁵ Abbink and Brandts (2008) and Argenton and Müller (2012) study the particular case of price competition with increasing marginal costs and no capacity constraints, where firms have to satisfy the whole demand at any price.

2.2 Sequential Choice of Quantities and Prices

A number of studies focus on duopolies in which decisions are made sequentially rather than simultaneously. Three important variations of sequential decision-making that have been studied experimentally are: (1) whether the order of decisions is exogenously fixed or emerges endogenously, (2) whether the decision of the first-mover is perfectly observable or not, and (3) whether the decision variable is quantity (with a homogeneous product) or price (with differentiated products).

Huck, Müller, and Normann (2001) and Kübler and Müller (2002) study the case with exogenous timing and perfect observability for quantity and price competition respectively. Huck, Müller, and Normann (2001) find some support for the notion of first-mover advantage in that leaders produce more than followers, but leaders tend to produce less than predicted and followers more than predicted in the subgame perfect equilibrium. They also find that total production is higher under simultaneous than under sequential decision-making. Kübler and Müller (2002) find that, consistent with the subgame perfect equilibrium, average prices set by leaders are higher than those set by followers and average profits of followers are higher than those of followers.

Huck and Müller (2000) and Morgan and Várdy (2004) study sequential quantity competition under exogenous timing and perfect observability. The work by Huck and Müller (2000) is directly motivated by the model of Bagwell (1995) where the pure-strategy equilibrium with first-mover advantage, i.e., the strategic benefit of committing oneself to a quantity level before the second mover, disappears completely if the action is only imperfectly observed by the second mover, with the slightest imperfection being enough to destroy the first-mover advantage. They test the Bagwell conjecture using three different levels of the quality of the signal that informs the second mover and find little support for it; first movers in experimental games do not lose their commitment power in the presence of noise. When the quality of signals is nearly perfect (99 percent) play almost completely converges to the Stackelberg outcome, whereas with lower signal quality the first-mover advantage remains but behavior is less well explained by the Stackelberg outcome.⁶ They also report that their data do not allow them to evaluate the noisy Stackelberg hypothesis of Van Damme and Hurkens (1997).

In the model by Várdy (2004), where the follower chooses whether to pay some cost to perfectly observe the leader's action, the value of commitment is completely shown in the unique pure-strategy subgame perfect equilibrium of this game; however, there exists a mixed-strategy subgame perfect equilibrium where the value of commitment is fully preserved. Morgan and Várdy (2004) find that in their experiments the value of commitment is largely preserved when the observation cost is small, while it is lost when the cost is large.

If the timing of moves is endogenous both simultaneous-move Cournot or Stackelberg outcomes may arise in equilibrium (Saloner, 1987; Hamilton and Slutsky, 1990; Ellingsen, 1995). For quantity competition the experimental results show that when both simultaneous move or Stackelberg outcomes can arise in equilibrium, Stackelberg leadership does not emerge easily (Huck, Müller, and Normann, 2002; Fonseca et al., 2005; Müller, 2006).⁷ For the case of

⁶ Georganas and Nagel (2011) study another environment where theory proposes that a small change leads to an "explosive" change, but the experimental data do not support this.

⁷ For an explanation of this result in terms of relative payoffs see Huck, Müller, and Normann (2001) and Santos-Pinto (2008).

price competition, Mago and Dechenaux (2009) find that quite a substantial degree of firm size asymmetry is needed for price leadership to emerge in posted-offer markets.

There are a number of studies pertaining to environments in which firms set both quantities and prices. A number of experiments have been inspired by the model of Kreps and Scheinkman (1983) in which firms first set production capacity and then, after observing capacities, set prices. For this model the equilibrium price is the same as in the Cournot equilibrium of quantity competition. Typically, in these experiments the stage game is repeated with the same group of sellers. A robust finding is that experience and learning are important determinants of outcomes. Sellers who have experience with repeatedly playing the stage game with the same partners choose capacities closer to the Cournot equilibrium quantity as compared to inexperienced sellers. Capacities chosen by inexperienced sellers are typically above the Cournot quantity, so more competitive (Davis, 1999; Muren, 2000; Anderhub et al., 2003; Goodwin and Mestelman, 2010; Le Coq and Sturluson, 2012; Hampton and Sherstyuk, 2012). Also, inexperienced sellers learn to set prices closer to the market clearing price level if they get the chance to learn the consequences of their prices, that is, if capacity choices are fixed for a number of rounds (Anderhub et al., 2003).⁸

2.3 Advance Production

There also some experimental studies in which firms set prices and at the same time pre-produce quantities incurring in production costs before they know how much they will sell. This is surely a relevant case, since many goods are produced before they are offered for sale at retail shops. Mestelman, Welland and Welland (1988) and Mestelman and Welland (1991) study advance production in a setting with discrete-unit step functions, where in the tradition of early market experiments cost and demand conditions are not provided as full information. The results of these two studies show that in comparison with posted-offer markets, advance production leads to lower efficiency levels and lower prices.

Brandts and Guillén (2007) study 50-round duopoly and triopoly market experiments in which firms decide repeatedly both on price and a pre-produced quantity of a completely perishable good. Each firm has the capacity to serve the whole market. The stage game does not have an equilibrium in pure strategies. Most markets evolve either to monopolies as a consequence of bankruptcies or to collusion at the monopolistic price. Evolution is faster in markets with two than in those with three firms. Therefore, over time average price is lower with three than with two. Over time consumer surplus is higher with three firms. However, due to overproduction, efficiency is lower in markets with three firms.

Davis (2013) studies the effects of advance production in a simple environment in which the mixed-strategy equilibria of the one-shot game of different treatment can be identified. In the equilibrium of the static game the presence of advance production raises prices and reduces consumer surplus but has no effect on prices. The experiment is implemented as an indefinitely repeated game with a number of certain rounds and a stopping rule used to terminate the sessions and applied to all rounds beyond the certain ones. On the basis of the minimum discount factor necessary to support an equilibrium at the highest possible price Davis (2013) predicts higher collusion without advance production. The experimental results show that in the absence of advance production prices and earnings are significantly above

⁸ Muren (2000) studies triopoly and the others duopoly markets.

and consumer surplus significantly below the Nash equilibrium production. In the presence of advance production, data are close to the Nash equilibrium of the static game. Indeed, transaction prices are lower with than without advance production whereas Nash equilibrium predicts the inverse order of prices.⁹

2.4 Supply Function Competition

Brandts, Pezanis-Christou, and Schram (2008) conducted the first lab experiment on supply function competition as introduced by Klemperer and Meyer (1989). They study the case of fixed groups and focus on two issues: the effect of the number of competitors and the impact of forward markets on prices. They find that triopolies yield lower prices than duopolies (notwithstanding the existence of multiple equilibria in both cases) and that the existence of forward markets also leads to lower prices as shown theoretically by Allaz and Vila (1993) for the case of quantity competition (see also Le Coq and Orzen, 2006). Their results also show that, both for duopolies and triopolies, prices are lower with supply functions than with quantity competition, but are above marginal cost, consistent with the notion that supply function competition yields prices between those of price and quantity competition.

Bolle et al. (2013) study supply function competition with a design, based on Holmberg (2007, 2008) with inelastic demand, a price cap and capacity constraints for which the stage game equilibrium is unique. The results show that overall the shape of the supply functions (increasing and mostly convex) is qualitatively in line with the equilibrium. They also find that in a market with symmetric firms there is some tacit collusion and that in a market with asymmetric firms, the larger firms bid more competitively than predicted, while the smaller firms provide less than equilibrium quantities.

Brandts, Reynolds, and Schram (2014) study supply function competition in an environment with pivotal suppliers, i.e., suppliers without which the total capacity of the market is not enough to satisfy the fixed demand. The results show that the more fundamental intuitions about the impact of pivotal power are supported by the data. Prices are higher when (some) firms are pivotal. The existence of aggregate excess capacity is not enough to guarantee competitive prices, in accordance both with the predictions of the supply equilibrium function (SFE) model based on divisible output and multi-unit auction (MUA) model based on discrete output units. Moreover, pivotal suppliers have a stronger impact on prices in a situation where productive capacity is symmetrically distributed than when the distribution is asymmetric. In other words having a number of suppliers with the same production capacity all being pivotal has a stronger upward effect on prices than some larger suppliers being pivotal, while smaller ones are not.

Overall, the results of the papers presented in Section 2 show that the equilibria of the relevant games are useful to organize the data from the experiments. However, there are also some important deviations that need to be better understood, like the deviations from equilibrium under Bertrand-Edgeworth competition.

⁹ Davis (2013) also studies the effects on one-period inventory carryover in an advance production setting. He finds that for this case the Nash equilibrium prescription is that prices will be lower than in the absence of advance production. However, the experimental results show that prices are significantly below those of posted prices with and without advance production. For other experiments on the effects of being able to carry over inventories see Mestelman and Welland (1988, 1991) and Reynolds (2000).

3 TACIT COLLUSION AND FACILITATING FACTORS¹⁰

Tacit collusion occurs when firms coordinate strategies in order to raise prices and profits without explicitly agreeing to do so. It is difficult to identify such conduct with field data because it is usually unknown what “non-collusive” prices and profits are. An advantage of the lab is that it is typically known what the static non-cooperative equilibrium is and behavior is termed collusive if aggregate outcomes are less competitive than in the static equilibrium (Holt, 1995). We now discuss some experiments that study conditions that favor collusion.¹¹

3.1 Supply and Demand Conditions

The prevalence of tacit collusion is strongly affected by the number of competitors. Overall, implicit coordination on a joint profit-maximizing price (or close to it) is frequently observed in markets with two sellers, rarely in markets with three sellers, and almost never in markets with four or more sellers.¹² This effect has been observed for quantity competition (Huck, Normann, and Oechssler, 2004), posted-offer markets (Davis, 2009; Ewing and Kruse, 2010; Fonseca and Normann, 2013), under Bertrand competition (Abdink and Brandts, 2005, 2008; Orzen, 2008), under advance production (Brandts and Guillén, 2007) and under supply function competition (Brandts et al., 2008).¹³

Another supply factor that has been shown to impact collusion is whether firms are symmetric or not. The evidence is mixed in this case. Cost asymmetries may hinder collusion (Mason, Phillips, and Nowell, 1992; Mason and Phillips, 1997) in Cournot duopolies, consistent with the standard view in anti-trust guidelines. Argenton and Müller (2012) extend the analysis to asymmetric Bertrand duopolies with convex costs, which have mixed strategy equilibria. They find, remarkably, that cost asymmetries facilitate collusion. The authors speculate that under asymmetry the low-cost firm acts as a price leader who is followed by the high-cost firm, making it easier to coordinate on a common price than under symmetry. Brandts et al. (2014) find that having symmetric firms all being pivotal leads to higher prices than having, in a setting with asymmetric firms, those of high capacity being pivotal.

Anderson et al. (2010) compare price-setting duopolies with substitute products (so that prices are strategic complements) to price-setting duopolies with complementary products (where prices are strategic substitutes). They find that, in the aggregate, the former markets are more collusive than the latter. This is remarkable, since it goes against the suggestion that games with strategic complements are not as competitive as games with strategic substitutes (Potters and Suetens, 2009). It is not entirely clear what drives this difference. Possibly, it is related to (the absolute value of) the slope of the best-response (BR) function, which affects the force of BR dynamics in pulling the outcome toward Nash equilibrium.¹⁴ A somewhat related issue is studied by Bruttel (2009b) when she compares price-setting duopolies with

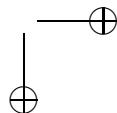
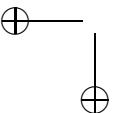
¹⁰ This section draws heavily on Potters and Suetens (2013).

¹¹ For a specific survey on collusion experiments see Haan, Schoonbeek, and Winkel (2009).

¹² Results from merger experiments are less straightforward, perhaps because mergers induce asymmetries (Huck, Konrad, and Müller, 2001; Davis, 2002; Davis and Wilson, 2005; Huck et al., 2007; Fonseca and Normann, 2008; Huck, 2009). See Lindqvist and Stenek (2005) for a study on endogenous merger formation, and Goette and Schmutzler (2009) for an excellent survey of the experimental literature on mergers.

¹³ It is interesting that in some studies prices increase as the number of firms increases Morgan et al. (2006a), but only so under random matching and if the underlying NE predicts this to occur. See also Orzen (2008).

¹⁴ For some evidence pointing in this direction see Cox and Walker (1998) and Chen and Gazzale (2004).



and without product differentiation. Her aim is to examine whether the analysis of the critical discount factor for the sustainability of cooperation in infinitely repeated games (Friedman, 1971) is relevant for tacit collusion in finitely repeated games as well. This seems to be the case indeed, as she finds less collusion with differentiated products than with homogeneous products.

Abbink and Brandts (2009) study collusion in Bertrand duopolies under dynamic demand conditions. In one treatment demand grows over time; in the mirror treatment demand declines over time. The results show that collusion is more frequent when demand grows than when it shrinks. The authors' conjecture is that the prospect of declining profits exerts a disciplining effect and discourages defection. Another demand factor that seems important in Bertrand markets is how demand is determined in the case that both firms offer the same price. Puzzello (2008) shows that collusion is easier if demand is shared equally than in the case where total demand is randomly allocated to either one of the two firms in the case of a tie. The effect is particularly strong when the price space is rather coarse.

3.2 Facilitating Institutions

An institution that seems of central importance for the scope for collusion is the possibility to monitor competitors' conduct, especially when price is a noisy signal of that conduct due to unobservable demand shocks. In line with the theoretical literature, which dates back to Stigler (1964), Feinberg and Snyder (2002) show that it is more difficult to collude when there is uncertainty about rivals' actions (see also Aoyagi and Fréchette, 2009). For Cournot triopolies, Offerman et al. (2002) find that collusion is more frequent when firms receive information about each competitor's quantity rather than about aggregate quantity. In the case that a market is not hospitable to collusion in the first place, matters are subtler.

Another facilitating institution that has drawn considerable interest is price-matching guarantees (PMGs). The predominant view in IO is that PMGs are anti-competitive since they reduce the incentives of firms to undercut their rivals. There are other perspectives though, such as the role of PMGs as credible price signals or as price discrimination devices. Field studies on the matter are rather scarce and inconclusive in all. Such a state of affairs calls for experiments and several researchers have picked up on that lately. The way PMGs are typically implemented in the lab is that a firm issues a price offer but that its effective price is equal to the lowest price offer in the market.

The experimental evidence suggests that such PMGs lead to higher prices, above the non-cooperative level. Fatás and Máñez (2007), for example, find that in a duopoly with differentiated goods prices are close to the collusive level if both firms implement a PMG, whereas prices are close to the non-cooperative level if neither firm implements a PMG. The potential loss of being undercut by a rival's price offer is entirely eliminated with a PMG in this setting. This anti-competitive effect holds both with homogeneous and with differentiated goods (Mago and Pate, 2009). Moreover, it does not seem to matter much whether the PMG is imposed exogenously or whether it is chosen by the firms themselves. In simple settings most subjects seem to realize very quickly that opting for a PMG is a profitable thing to do (Fatás and Máñez, 2007), although in more complex settings this seems less obvious (Deck and Wilson, 2003).

It appears that the collusion-facilitating effect of PMGs is robust to products being homogeneous rather than heterogeneous (Dugar, 2007; Mago and Pate, 2009), to the market

having two, three or four firms (Deck and Wilson, 2003; Dugar, 2007), to firms having asymmetric costs (Mago and Pate, 2009), and a design with strangers or partners matching (Dugar, 2007). There are other factors though that can substantially reduce the collusive effect of PMGs. One is the presence of hassle costs, due to which it is costly for buyers to effectuate a PMG (Dugar and Sorensen, 2006). Another is the use of a more aggressive price-beating guarantee that ensures that a lower price of a competitor is not matched but undercut (Fatás et al., 2005, 2013). Using a relatively elaborate design with both human sellers and buyers Yuan and Krishna (2011) show that when buyers need to search for price information and informed buyers have more elastic demand than uninformed buyers, PMGs may even be pro-competitive as they increase buyers' incentives to search. So, experiments have generated many useful insights, but the jury is still out on whether PMGs are predominantly collusive.

Finally, one recent paper studies capacity coordination. In the case of an unexpected negative demand shock an argument in favor of capacity coordination is that it will prevent the duplication of fixed but avoidable costs. The risk, however, is that it will facilitate tacit price collusion. Hampton and Sherstyuk (2012) implement a repeated Kreps-Scheinkman two-stage capacity and price-setting game in which halfway there is a demand shock and they compare treatments with and without explicit capacity coordination. They find, first, that explicit capacity coordination is not necessary for a quick adjustment of capacities after the shock, and, second, that explicit capacity coordination has a pronounced effect on collusion. The net effect of capacity coordination on welfare is clearly negative.

4 PRODUCT DIFFERENTIATION

In this section we review experiments that study horizontal (spatial) product differentiation and we focus on settings in which differentiation is endogenous. Brown-Kruse, Cronshaw, and Schenk (1993) is the first experimental study on location choice in a classic Hotelling framework. They implemented markets with two sellers who had to choose a location along a road represented by the interval $[0, 100]$. The unit price was fixed and the same for the two sellers. Consumers were located uniformly along the road. Demand by a consumer was equal to $10 - p - 0.1d$, where p is the unit price (set at $p = 0.53$) and d is the distance of the consumer to the seller chosen. Marginal costs were equal to 0.5; fixed costs were equal to 10. The profit of seller i was equal to $0.03Q_i - 10$, where Q_i is the sum of demands for consumers who buy from seller i . Buyers were simulated and always chose the closest seller. In the case of a tie demand was split.

At the beginning of an experimental session, subjects were matched in pairs who remained together throughout the experiment. In each period, subjects had to simultaneously choose their location. After that they were informed about the quantity they sold, their market share, and the profit for the period. The length of the game was determined randomly. At the end of each period, there was a $7/8$ chance that the game would continue and a $1/8$ chance that the game would end. Theoretically, this procedure is equivalent to having an infinitely repeated game with a discount factor of $7/8$.

With this parameterization, the unique symmetric Nash equilibrium of the stage game is for both sellers to locate in the middle. This equilibrium is Pareto-dominated by the collusive outcome in which the sellers locate at the quartiles (25 and 75). Both outcomes

(50, 50) and (25, 75) can be sustained as equilibria of the indefinitely repeated game. It is not straightforward to predict which of these outcomes will have the strongest drawing power. This is what makes the experimental design particularly interesting. Another attractive feature of the set-up is that there were two treatments: one in which no communication was allowed between sellers, and one treatment in which sellers in the same market were allowed to engage in anonymous non-binding free-form communication.

The experiment used 48 student subjects drawn from undergraduate microeconomic classes. The duration of the markets varied from four to 15 periods. The results indicate that market outcomes strongly depend on the possibility to communicate. In the treatment without communication, sellers predominantly located near the center. The majority of location choices were within the range 48–52, with the modal choice being at 50. Even though there were a few unilateral attempts to move away from the center, these attempts were never successful. The results are dramatically different when sellers were allowed to communicate. Now, the predominant outcome was for one seller to locate at 25 and the other at 75. These results indicate, once more, the strong effect of communication on repeated market interaction. Without communication sellers clustered near the center and observed the principle of minimum differentiation. With communication, even though anonymous and non-binding, sellers managed to coordinate on the payoff-dominant quartile equilibrium.

Later studies extend the experiment by Brown-Kruse et al. (1993) in several directions. Brown-Kruse and Schenk (2000) vary the location game along two dimensions. One made the game simpler by allowing each seller to choose from only two locations along the road, rather than any location. The hypothesis that a simpler game would lead to more cooperation in the absence of communication, was supported by the results. The other variation was that, apart from a uniform distribution, a unimodal distribution with consumers concentrated near the center was also implemented, as well as a bimodal distribution with consumers concentrated near the endpoints of the product space. The hypothesis was that as more (less) consumers are concentrated near the center, coordination on the Pareto-dominant quartile equilibrium should be more difficult (easier). The experimental results lend little support to this hypothesis. Without communication, sellers had difficulty in coordinating on the joint profit maximum, irrespective of the distribution of consumers. With communication, sellers managed to coordinate well even in case when the consumers were concentrated in the center. The increased incentives to defect were not able to upset cooperation.

Collins and Sherstyuk (2000) examine a three-seller version of the Hotelling model, in which the unique symmetric equilibrium (under risk neutrality) is for each seller to randomly choose a location in the interval [25, 75]. In the experiment, markets operate for a fixed number of periods and without communication. The experimental results confirm that locations near the edges are rare. Location choices were not uniform but bimodal with peaks in between the center and the quartiles. Huck, Müller, and Vriend (2002) implement a four-player location game, in which the pure-strategy Nash equilibria consist of two sellers locating at 25 and two at 75. The results display three clusters of location choices. Besides the two equilibrium locations, the center was frequently chosen and this frequency increased over time. The authors attribute this to best reply dynamics that drag the players toward the center.

Barreda-Tarazona et al. (2011) complement location choice with an endogenous price setting. In addition to a demand effect, location choice now involves a strategic price effect. In the experiment two sellers first simultaneously choose their locations and then

choose prices. The results display a strong tendency to locate in the center. In a majority of cases there was little or no differentiation. Average prices display a positive relationship with differentiation, although the relationship was a bit weaker than predicted by non-cooperative equilibrium. Finally, the incidence of collusion, in either locations or prices, was rather low.

Product differentiation may ease competition whether it is “real” or “spurious.” Firms may thus have incentives to make products look more different than they really are and to make it harder for consumers to compare products in order to reduce the price elasticity of demand. Obviously, this only works if consumers are boundedly rational and if firms anticipate that they are. Kalayci and Potters (2011) present a model and conduct an experiment in which sellers can raise the number of attributes of their products. Adding attributes does not change the value of products for buyers, it only makes it more difficult for buyers to assess and compare the value of different products. The experimental results show that when the number of attributes is higher buyers make more suboptimal choices and sellers charge higher prices. The intuition that it may be profitable to obfuscate consumers appears to be very compelling even to inexperienced sellers.

Kalayci (2015a) investigates whether the incentive to obfuscate extends to prices. In the experiment, sellers choose both the price level and the complexity of the (multi-part) price structure. The evidence is somewhat mixed. Complex prices on average are high prices, as predicted by the model, but this holds only if the buyer strategies are simulated and not if the buyers are human subjects. Kalayci (2015b) and Normann and Wenzel (2015) vary the number of competitors. Kalayci (2015b) finds that obfuscation is unrelated to numbers, while Normann and Wenzel find that add-ons are shrouded more in a duopoly than in a quadropoly. Gu and Wenzel (2015) implement a duopoly model in which firms have asymmetric incentives to obfuscate consumers. In line with the model’s predictions the experimental results illustrate that creating a level playing field may not always be beneficial as it may increase the incentives of firms to obfuscate.

5 EXPERIENCE AND CREDENCE GOODS

Asymmetric information and moral hazard may be serious impediments to market efficiency. If consumers cannot ascertain the quality of a good or service before purchase, sellers may have little incentive to provide high quality. That this moral hazard problem is “real” has already been demonstrated in experiments conducted in the 1980s and early 1990s (DeJong et al., 1985; Lynch et al., 1986; Holt and Sherman, 1990). For example, Lynch et al. (1986) implemented markets for experience goods in which each seller had to choose whether to produce low-quality or high-quality products. The exchange of high quality would be efficient. Sellers could announce the quality they offered but buyers could not ascertain quality before purchase. The experiments predominantly led to outcomes in which mainly low quality was produced. High-quality products could not command a high enough price premium to make up for the higher production costs.

A range of remedies has been suggested for this problem, and several of these have been put to experimental tests. An effective one is a disclosure rule that prevents sellers from overstating the quality of their product (e.g., Lynch et al., 1986; Forsythe et al., 1999). The effectiveness of this remedy is perhaps not so surprising from a theoretical perspective.

Also one may wonder how realistic the implementation of these perfect remedies is. More interesting and challenging are the effectiveness of remedies such as signaling, reputation formation, and competition.¹⁵

Whether costly signaling can prevent lemons outcomes is investigated by Miller and Plott (1985). Sellers were exogenously endowed with low- or high-quality products. Quality was unobservable to buyers, but sellers could send observable signals (think of warranties) that were more costly for low-quality than for high-quality sellers. Exchange took place in a double auction market in which sellers and buyers could submit price offers and asks for products with specific signal levels. This set-up allows for both separating and pooling equilibria. The experimental results indicate that some markets exhibit (partial) pooling. Quality separating occurs in all markets in which the marginal costs of signaling high quality is relatively low. A typical time pattern is for signaling to start at inefficiently high levels – to establish clear separation – and then to come down to an efficient level that just deters low qualities from sending the same signal.

Davis and Holt (1994a) examine the force of repeated interaction in a three-player game that resembles a market for an experience good. In the game a buyer chooses between one of two sellers, and the selected seller chooses whether to provide low quality or high quality. Upon experiencing this quality level, the buyer then chooses between one of the two sellers in the next period. This game has multiple equilibria, including an equilibrium in which both sellers always choose high quality except for the terminal period. A buyer strategy that induces such an outcome is to stay with the same seller if and only if the experienced quality is high. The results indicate that high quality was delivered at a rate of 0.63 in the case that the game was repeated for ten periods, but only at a rate of 0.25 in the case that the game was played once or repeated only twice.

Huck, Lünser, and Tyran (2012) implement a trust game that can be interpreted as a market for experience goods with fixed prices and endogenous quality. A market consists of four buyers and four sellers. A buyer can buy from at most one seller, and a seller can supply up to four buyers. The experimental markets run for 30 periods and information and competition vary along two dimensions. Under private information, buyers learn only the quality delivered by the seller they bought from; under full information buyers learn the qualities delivered by all sellers. With competition, buyers can decide from which seller they want to buy; without competition, buyers are randomly assigned to a seller and then decide to buy or not. The experimental results display a strong effect of competition. Efficiency goes up from 36 percent without competition to about 80 percent with competition. The fact that buyers can choose whom to buy from, disciplines the sellers and increases the willingness of buyers to buy. Still, the fact that sellers can form a reputation is important even if there is no competition. In a control treatment in which repeat purchases are excluded, efficiency is as low as 8 percent. Whether reputational information is private or full does not matter so much. Buyers seem to focus mainly on their own experiences and not so much on those of other buyers. In a follow-up paper, Huck, Lünser, and Tyran (2016) introduce price competition in the market for experience goods. This turns out to lead to a strong negative impact on efficiency. The reason is that buyers no longer focus exclusively on sellers' reputations but are also attracted by low prices. These, however, are typically offered by less reputable sellers.

¹⁵ Another remedy with non-trivial effects is (partial) transparency, which reduces or eliminates the informational asymmetry between buyers and sellers (Henze, Schuett, and Sluijs, 2015).

Markets for credence goods (such as car repairs or medical treatments) are even more intricate than those for experience goods. The quality of credence goods cannot be assessed by buyers even after purchase. A buyer may then face three types of problems: overtreatment (unnecessary repairs are made), undertreatment (necessary repairs were not made), and overcharging (fictitious repairs are charged for). In an unusually rich experiment, Dulleck, Kerschbamer, and Sutter (2011) implement a setting in which sellers first set prices for low quality and for high quality, buyers decide the buy from a seller or not, nature reveals to the seller whether the buyer needs low or high quality, the seller decides whether to supply low quality or high quality and the whether to charge the low or the high price. In a $2 \times 2 \times 2 \times 2$ design, Dulleck et al. (2011) examine the force of four potential remedies: competition (buyers can choose between sellers), reputation formation (buyers can identify sellers), liability (sellers cannot undertreat), and verifiability (sellers cannot overcharge). The equilibrium predicts that verifiability and liability are equally effective remedies against the moral hazard problem, but that reputation and competition are not. The experimental results indicate that market efficiency is very low (18 percent) when none of the potential remedies are operative; market failure is a real threat to credence goods. The introduction of competition does not improve efficiency at all (13 percent), while the possibility to form a reputation does somewhat better (27 percent). Reputation formation seems less effective for credence goods than for experience goods, possibly because buyers cannot even observe quality *ex post*. Another main result is that liability is an effective remedy against market failure (efficiency 84 percent) while verifiability is not at all (efficiency 16 percent). One reason is that the effectiveness of verifiability depends on sellers charging the right prices, whereas the effect of liability on undertreatment is direct. Liability does, however, also lead to much higher prices (which may be illustrative for healthcare markets and medicines).

All of the experimental papers discussed here document a lower degree of opportunistic behavior than could be expected if players were unboundedly selfish. In line with the evidence from hundreds of experimental studies in related fields it appears that social preferences and belief-based motivations may exercise a check on moral hazard. Targeted evidence is provided in Beck et al. (2013) who show that the possibility of sellers making non-binding promises may curtail opportunistic behavior. In line with a model of guilt aversion (Battigalli and Dufwenberg, 2007) many sellers do not wish to betray the trust that buyers place on them after being promised fair treatment. On a different note, Kerschbamer, Sutter, and Dulleck (2015) delve deeper into the finding that verifiability is a relatively ineffective remedy and find that this may be due to the presence of social preferences. As social preferences differ across sellers (being positive for some and negative for others) this renders it hard for prices to make sellers indifferent between providing low and high quality, which is necessary for verifiability to ensure efficient quality provision. So, here we have a case in which social preferences may interact in intricate ways with the effectiveness of institutions.

In sum, the experimental evidence indicates that asymmetric information and moral hazard constitute “real” market failures. Strong institutional interventions such as an anti-fraud rule that prohibits false quality claims or a liability rule that bans underprovision are effective remedies. The question though is how realistic such interventions are. More subtle mechanisms such as costly signaling and reputation can also generate trust and trustworthiness, but only, it seems, if their informational value can be assessed easily and does not need to compete with other attributes (like prices).

6 STRATEGIC BEHAVIOR

6.1 Entry Deterrence

Cooper, Garvin, and Kagel (1997a, 1997b) put a version of the limit-pricing model by Milgrom and Roberts (1982) to an experimental test.¹⁶ A monopolist faces a potential entrant and can try to set a low price in order to deter entry. The monopolist can be of two types, low cost or high cost. For the entrant it is profitable to enter if and only if monopolist has high costs. Different games with different sets of equilibria are implemented. The experimental results show that limit pricing takes time to develop. In the early rounds, monopolists ignore the threat of entry. This induces high entry rates for the high-cost monopolist, who then attempts to pool with the low-cost monopolist. When pooling is not an equilibrium, because the prior beliefs induce entry, this forces the low-cost monopolist to choose a lower price in order to separate from the high-cost monopolist. The experimental results show that play develops over time and is history dependent. An adaptive learning model captures behavior well.

Müller, Spiegel, and Yehezkel (2009) implement the entry deterrence model by Bagwell and Ramey (1991) in which there are two incumbents who can distort prices in order to signal to a potential entrant that cost conditions are unfavorable. There are multiple equilibria, both pooling and separating. The experimental results reveal strong support for the “full information equilibrium.” Incumbents choose prices as if there were no asymmetry of information, and are able to deter entry when costs are high without having to distort prices.

Investment in capacity is another means by which incumbents can try to deter entry. A simple Dixit-style model was first investigated experimentally by Mason and Nowell (1992). In the unique subgame perfect equilibrium an incumbent chooses an output level that is just large enough to make entry by a potential entrant unprofitable. The results indicate that many subjects play the subgame perfect equilibrium and the rate at which they do increases with learning. Still, a fraction of the incumbents chooses not to deter entry and a portion of the entrants keeps entering even when this entails negative profits. One explanation for the latter effect is that entrants try to build a reputation for toughness.

Brandts, Cabrales, and Charness (2007) investigate a model based on Bagwell and Ramey (1991) in which not only the incumbent can invest in capacity (in the first stage) but also the potential entrant can precommit (in the second stage). This renders two subgame perfect equilibria, where either the incumbent or the entrant becomes the monopolist (in the third stage). A forward induction argument – an entrant who precommits can do so profitably only if she believes to be the monopolist – would favor the equilibrium in which the entrant precommits and the incumbent does not. The experimental results do not show much support for this equilibrium. Precommitment by either player is limited, and it is three times more likely that the incumbent becomes the monopolist than the entrant. This points toward a strategic first-mover advantage that is not tied to the force of forward induction.

¹⁶ For experimental tests of the related reputation model of Kreps and Wilson (1982) see Jung, Kagel, and Levin (1994).

6.2 Research and Development

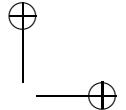
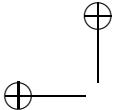
One of the first papers to study R&D in the lab is Isaac and Reynolds (1988). They implement a stochastic invention model in which firms simultaneously decide how much to spend on R&D and the probability of an innovative success increases (concavely) in the R&D spending. Using a 2×2 design, the experiment reveals support for two comparative statics predictions. First, individual investment levels decrease with the number of firms, while total investments increase. Second, full appropriability of innovation (i.e., the absence of spillovers) increases investment levels relative to partial appropriability. Moreover, with full appropriability investment levels exceed the socially optimal level, although not as much as equilibrium predicts.

Several other studies have studied the effect of spillovers on R&D investment and innovation. Suetens (2005, 2006) implements a simplified version of the (non-stochastic) model of D'Aspremont and Jacquemin (1988) in which firms first decide on cost-reducing R&D investments and then compete on a product market. She finds strong support for the prediction that investment levels decrease with spillovers. Without spillovers, investments are typically lower than the cooperative joint profit maximizing, and investments are typically above the cooperative level without spillovers and below the cooperative level with spillovers (see also Halbheer et al., 2009). Moreover, the effect of communication varies with spillovers. When there are no spillovers, cheap talk hardly has an effect on outcomes, but with spillovers cheap talk increases investment levels close to cooperative levels.

Another recurring topic is the effect of market structure on innovation. Darai, Sacco, and Schmutzler (2010) implement a 2×2 design in which they vary the number of firms from two to four and also compare quantity and price competition. The experimental results show that moving from two to four firms reduces cost-reducing R&D investments, as predicted by equilibrium. Moving from quantity to price competition increases investments, which is in line with equilibrium for duopolies but contrary to equilibrium for the four-firm case. Sacco and Schmutzler (2011) examine the effect of competition by varying the degree of product differentiation. They report substantial support for the predicted U-shaped relationship. When firms compete in quantities investments in cost reduction are lowest at intermediate levels of product differentiation, and higher when products are either homogeneous or non-substitutable.

A final theme in several experiments is the effect of asymmetry in innovation. A dynamic stochastic patent race, based on Harris and Vickers (1987), was implemented by Zizzo (2002). The experiment provides little support for the model's prediction that leaders should invest more than followers. Limited support for the theoretical predictions in a dynamic patent race model is also reported in Breitmoser et al. (2010). More support for the prediction that asymmetries tend to be reinforced by innovation incentives is reported in Halbheer et al. (2009). They implement a deterministic model of cost-reducing innovation. Since low-cost firms have larger market shares they also have a higher incentive to reduce marginal cost even further. Such a self-reinforcing effect of market dominance is strongly supported by the experimental results. A recent study by Aghion et al. (2014), implements the step-by-step model by Aghion et al. (2001) in the lab, and finds support for the prediction that the effect of competition on innovation depends on the degree of asymmetry between firms.

Overall, it seems that the experimental evidence provides rather broad support for deterministic models that relate innovation incentives to appropriability and market structure. Somewhat weaker support is found for patent race models that have a strong stochastic component.



7 CONCLUSION

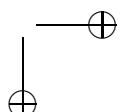
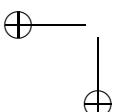
We have surveyed a broad and diverse set of experiments in industrial organization. We cannot claim that we have been complete (areas not covered include consumer search, antitrust, entry games, strategic delegation); we only hope that we have been able to illustrate that experimentation in IO is an active and rich research area with many important and interesting results. We will not attempt to summarize these results here, but just offer some final remarks.

We believe it is fair to say that equilibrium is a powerful tool in predicting comparative statics. When equilibrium indicates that an increase in X will increase the occurrence of Y, such predictions are often borne out by the experimental data. One can think of how numbers affect prices, how spillovers affect innovation, or how liability affects moral hazard. The size of the effects is a different matter. This is usually predicted much less accurately. This is partly due to the fact that best responses depend only on the sign of payoff differences. As soon as one action is just a bit better than other actions, it is predicted to be chosen with certainty. Behavioral responses, however, typically also depend on the size of payoff differences; the larger the payoff difference between a strategy and other strategies, the more likely it will be chosen, as in the quantal-response equilibrium of McKelvey and Palfrey (1995).

This connects to the more general issue of whether behavior in industrial organization experiments can be fruitfully studied using some of the other behavioral theory models that to a large extent were the outgrowth of experimental work like models involving cognitive hierarchies or relative payoff considerations and efficiency concerns.

Another reason why behavior may differ from standard predictions is that players sometimes do better than equilibrium would suggest. This is the case, for example, when players manage to collude in finitely repeated games, or when they succeed in circumventing moral hazard problems. When equilibrium is inefficient and it is obvious that there are other outcomes that are better for everyone, then players are sometimes able to attain those outcomes. This holds in particular for repeated two-player interactions, or when players can communicate to each other. Even when standard theory suggests that talk is cheap, communication can have powerful effects. This is one more area where promising new theories are beginning to develop and where this development is greatly facilitated by a close interaction between modeling and experimenting (e.g., Charness and Dufwenberg, 2006).

Any study, whether theoretical or empirical, raises questions of generalizability. In the area of experimental industrial organization one may wonder whether conclusions about the predictive value of a model are similar if one conducts experiments with professionals from the field rather than with students. Fréchette (2015) presents a systematic exploration and concludes that overall, studies with student and with professionals lead to similar conclusions with respect to how the behavior of subjects conforms to the comparative static predictions of a theory. Another issue is that decisions in firms are often taken by groups of individuals and that decision-makers act as agents on behalf of a principal. Some studies find that the behavior of teams and agents is closer to standard equilibrium predictions, while others do not find behavioral differences between teams and individuals. Engel (2010) provides a comprehensive treatment of these issues, and concludes that “in many respects, collective and corporate actors suffer from the same biases as individuals” (p. 463). We would like to reiterate though that laboratory experiments can be replicated. The robustness of the results can be tested and generalizability can be explored using a different subject pool or implementing changes in design and procedure. If results survive such scrutiny one can become progressively more



confident in their validity. If results do not generalize, one can trace the factors that are responsible for this and adapt models if necessary.

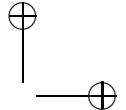
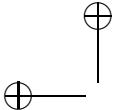
Finally, let us point out two areas in industrial organization where we hope that experimental economics will make significant contributions in the near future. One emerging field is behavioral industrial organization. This mostly theoretical literature studies whether and how firms can take advantage of the behavioral biases of consumers (see Spiegler, 2011; Grubb, 2015). It has been inspired, at least partially, by experimental work documenting phenomena like loss aversion, over-confidence and suboptimal search. It is an interesting twist that these theoretical models can be brought back into the lab. The challenge for experimenters is to bring these models into the lab, to create an environment in which the relevant behavioral biases of consumers arise and analyze how observed behavior relates to the equilibrium of the behavioral model.¹⁷

Another promising area is the use of experiments to design institutions, which can be considered for adoption by policy-makers. An interesting example of such work is Goeree et al. (2010), which studies different institutional arrangements for pollution permits. Apesteguia, Dufwenberg, and Selten (2007) and Bigoni et al. (2012) conduct experiments to inform the design of leniency schemes and anti-trust fines to combat collusion. More research along these lines would certainly be of great interest.

REFERENCES

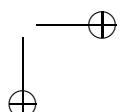
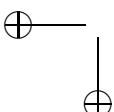
- Abbink, K. and Brandts, J. (2005). Price competition under cost uncertainty: a laboratory analysis. *Economic Inquiry*, 43(3), 636–648.
- Abbink, K. and Brandts, J. (2008). Pricing in Bertrand competition with increasing marginal costs. *Games and Economic Behavior*, 63(1), 1–31.
- Abbink, K. and Brandts, J. (2009). Collusion in growing and shrinking markets: empirical evidence from experimental duopolies. In: J. Hinloopen and H.T. Normann (eds), *Experiments and Competition Policy*. Cambridge, UK: Cambridge University Press.
- Abbink, K. and Brandts, J. (2010). Drei Oligopolexperimente. In: *The Selten School of Behavioral Economics: A Collection of Essays in Honor of Reinhard Selten*. Berlin/Heidelberg: Springer.
- Aghion, P., Bechtold, S., Cassar, L., and Herz, H. (2014). The causal effects of competition on innovation: experimental evidence, *NBER Working Paper* 19987.
- Aghion, P., Harris, C., Howitt, P., and Vickers, J. (2001). Competition, imitation and growth with step-by-step innovation. *Review of Economic Studies*, 68(3), 467–492.
- Allaz, B. and Vila, J.-L. (1993). Cournot competition, forward markets and efficiency. *Journal of Economic Theory*, 59, 1–16.
- Anderhub, V., Güth, W., Kamecke, U., and Normann, H.T. (2003). Capacity choices and price competition in experimental markets. *Experimental Economics*, 6(1), 27–52.
- Anderson, L.R., Freeborn, B.A., and Holt, C.A. (2010). Tacit collusion in price-setting duopoly markets: experimental evidence with complements and substitutes. *Southern Economic Journal*, 76(3), 577–591.
- Aoyagi, M. and Fréchette, G. (2009). Collusion as public monitoring becomes noisy: experimental evidence. *Journal of Economic theory*, 144(3), 1135–1165.
- Apesteguia, J., Dufwenberg, M., and Selten, R. (2007). Blowing the whistle. *Economic Theory*, 31(1), 143–166.
- Argenton, C. and Müller, W. (2012). Collusion in experimental Bertrand duopolies with convex costs: the role of cost asymmetry. *International Journal of Industrial Organization*, 30(6), 508–517.
- Aumann, R.J. and Shapley, L.S. (1994). *Long-term Competition – A Game-theoretic Analysis*. New York: Springer.
- Bagwell, K. (1995). Commitment and observability in games. *Games and Economic Behavior*, 8(2), 271–280.
- Bagwell, K. and Ramey, G. (1991). Oligopoly limit pricing. *The RAND Journal of Economics*, 22(2), 155–172.

¹⁷ Some recent contributions in this line by Kalayci and Potters (2011), Kalayci (2015a, 2015b), Normann and Wenzel (2015) and Gu and Wenzel (2015) have been discussed above.

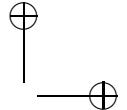
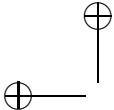


470 *Handbook of game theory and industrial organization: applications*

- Barreda-Tarazona, I., García-Gallego, A., and Georgantzís, N. et al. (2011). An experiment on spatial competition with endogenous pricing. *International Journal of Industrial Organization*, 29(1), 74–83.
- Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2), 170–176.
- Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2013). Shaping beliefs in experimental markets for expert services: guilt aversion and the impact of promises and money-burning options. *Games and Economic Behavior*, 81, 145–164.
- Bigoni, M., Fridolfsson, S.O., Le Coq, C., and Spagnolo, G. (2012). Fines, leniency, and rewards in antitrust. *The RAND Journal of Economics*, 43(2), 368–390.
- Bolle, F., Grimm, V., Ockenfels, A., and Del Pozo, X. (2013). An experiment on supply function competition. *European Economic Review*, 63, 170–185.
- Bosch-Domènec, A. and Vriend, N.J. (2003). Imitation of successful behaviour in Cournot markets. *The Economic Journal*, 113(487), 495–524.
- Brandts, J., and Guillén, P. (2007). Collusion and fights in an experiment with price-setting firms and advance production. *The Journal of Industrial Economics*, 55(3), 453–473.
- Brandts, J., Cabrales, A., and Charness, G. (2007). Forward induction and entry deterrence: an experiment. *Economic Theory*, 33(1), 183–209.
- Brandts, J., Pezanis-Christou, P., and Schram, A. (2008). Competition with forward contracts: a laboratory analysis motivated by electricity market design. *The Economic Journal*, 118(525), 192–214.
- Brandts, J., Reynolds, S.S., and Schram, A. (2014). Pivotal suppliers and market power in experimental supply function competition. *The Economic Journal*, 124, 887–916.
- Breitmoser, Y., Tan, J.H., and Zizzo, D.J. (2010). Understanding perpetual R&D races. *Economic Theory*, 44(3), 445–467.
- Brown-Kruse, J. and Schenk, D.J. (2000). Location, cooperation and communication: an experimental examination. *International Journal of Industrial Organization*, 18(1), 59–80.
- Brown-Kruse, J., Cronshaw, M.B., and Schenk, D.J. (1993). Theory and experiments on spatial competition. *Economic Inquiry*, 31(1), 139–165.
- Bruttel, L.V. (2009a). Group dynamics in experimental studies – the Bertrand Paradox revisited. *Journal of Economic Behavior and Organization*, 69(1), 51–63.
- Bruttel, L.V. (2009b). The critical discount factor as a measure for cartel stability? *Journal of Economics*, 96(2), 113–136.
- Cason, T.N. and Friedman, D. (2003). Buyer search and price dispersion: a laboratory study. *Journal of Economic Theory*, 112(2), 232–260.
- Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6), 1579–1601.
- Chen, Y. and Gazzale, R. (2004). When does learning in games generate convergence to Nash equilibria? The role of supermodularity in an experimental setting. *The American Economic Review*, 94, 1505–1535.
- Collins, R. and Sherstyuk, K. (2000). Spatial competition with three firms: an experimental study. *Economic Inquiry*, 38(1), 73–94.
- Cooper, D., Garvin, S., and Kagel, J.H. (1997a). Signalling and adaptive learning in an entry limit pricing game. *The RAND Journal of Economics*, 28(4), 662–683.
- Cooper, D.J., Garvin, S., and Kagel, J.H. (1997b). Adaptive learning vs. equilibrium refinements in an entry limit pricing game. *The Economic Journal*, 107(442), 553–575.
- Cox, J. and Walker, M. (1998). Learning to play Cournot duopoly strategies. *Journal of Economic Behavior and Organization*, 36(2), 141–161.
- Darai, D., Sacco, D., and Schmutzler, A. (2010). Competition and innovation: an experimental investigation. *Experimental Economics*, 13(4), 439–460.
- D'Aspremont, C. and Jacquemin, A. (1988). Cooperative and noncooperative R&D in duopoly with spillovers. *The American Economic Review*, 78(5), 1133–1137.
- Davis, D.D. (1999). Advance production and Cournot outcomes: an experimental investigation. *Journal of Economic Behavior and Organization*, 40(1), 59–79.
- Davis, D.D. (2002). Strategic interactions, market information and predicting the effects of mergers in differentiated product markets. *International Journal of Industrial Organization*, 20(9), 1277–1312.
- Davis, D.D. (2009). Pure numbers effects, market power, and tacit collusion in posted offer markets. *Journal of Economic Behavior and Organization*, 72(1), 475–488.
- Davis, D.D. (2011). Behavioral convergence properties of Cournot and Bertrand markets: an experimental analysis. *Journal of Economic Behavior and Organization*, 80(3), 443–458.
- Davis, D.D. (2013). Advance production, inventories, and market power: an experimental investigation. *Economic Inquiry*, 51(1), 941–958.
- Davis, D.D. and Holt, C.A. (1994a). Equilibrium cooperation in three-person, choice-of-partner games. *Games and Economic Behavior*, 7(1), 39–53.
- Davis, D.D. and Holt, C.A. (1994b). Market power and mergers in laboratory markets with posted prices. *The RAND Journal of Economics*, 25, 467–487.

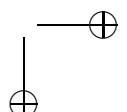
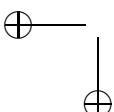


- Davis, D.D. and Holt, C.A. (1996). Consumer search costs and market performance. *Economic Inquiry*, 34(1), 133–151.
- Davis, D.D. and Wilson, B.J. (2005). Differentiated product competition and the antitrust logit model: an experimental analysis. *Journal of Economic Behavior and Organization*, 57(1), 89–113.
- Deck, C.A. and Wilson, B.J. (2003). Automated pricing rules in electronic posted offer markets. *Economic Inquiry*, 41(2), 208–223.
- DeJong, D.V., Forsythe, R., and Lundholm, R.J. (1985). Ripoffs, lemons, and reputation formation in agency relationships: a laboratory market study. *The Journal of Finance*, 40(3), 809–820.
- Dufwenberg, M., and Gneezy, U. (2000). Price competition and market concentration: an experimental study. *International Journal of Industrial Organization*, 18(1), 7–22.
- Dugar, S. (2007). Price-matching guarantees and equilibrium selection in a homogeneous product market: an experimental study. *Review of Industrial Organization*, 30(2), 107–119.
- Dugar, S. and Sorensen, T. (2006). Hassle costs, price-matching guarantees and price competition: an experiment. *Review of Industrial Organization*, 28(4), 359–378.
- Dulleck, U., Kerschbamer, R., and Sutter, M. (2011). The economics of credence goods: an experiment on the role of liability, verifiability, reputation, and competition. *American Economic Review*, 101, 526–555.
- Durham, Y., McCabe, K., and Olson, M.A. et al. (2004). Oligopoly competition in fixed cost environments. *International Journal of Industrial Organization*, 22(2), 147–162.
- Ellingsen, T. (1995). On flexibility in oligopoly. *Economics Letters*, 48(1), 83–89.
- Engel, C. (2007). How much collusion? A meta-analysis of oligopoly experiments. *Journal of Competition Law and Economics*, 3(4), 491–549.
- Engel, C. (2010). The behaviour of corporate actors: how much can we learn from the experimental literature? *Journal of Institutional Economics*, 6(04), 445–475.
- Ewing, B.T. and Kruse, J.B. (2010). An experimental examination of market concentration and capacity effects on price competition. *Journal of Business Valuation and Economic Loss Analysis*, 5(1).
- Falk, A. and Heckman, J.J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326, 535–538.
- Fatás, E. and Máñez, J.A. (2007). Are low-price promises collusion guarantees? An experimental test of price matching policies. *Spanish Economic Review*, 9(1), 59–77.
- Fatás, E., Georgantzís, N., Máñez, J.A., and Sabater-Grande, G. (2005). Pro-competitive price beating guarantees: experimental evidence. *Review of Industrial Organization*, 26(1), 115–136.
- Fatás, E., Georgantzís, N., Máñez, J.A., and Sabater-Grande, G. (2013). Experimental duopolies under price guarantees. *Applied Economics*, 45(1), 15–35.
- Feinberg, R. and Snyder, C. (2002). Collusion with secret price cuts: an experimental investigation. *Economics Bulletin*, 3(6), 1–11.
- Fonseca, M.A. and Normann, H.T. (2008). Mergers, asymmetries and collusion: experimental evidence. *The Economic Journal*, 118(527), 387–400.
- Fonseca, M.A. and Normann, H.T. (2013). Excess capacity and pricing in Bertrand-Edgeworth markets: experimental evidence. *Journal of Institutional and Theoretical Economics*, 169(2), 199–228.
- Fonseca, M.A., Huck, S., and Normann, H.T. (2005). Playing Cournot although they shouldn't. *Economic Theory*, 25(3), 669–677.
- Forsythe, R., Lundholm, R., and Rietz, T. (1999). Cheap talk, fraud, and adverse selection in financial markets: some experimental evidence. *Review of Financial Studies*, 12(3), 481–518.
- Fréchette, G. (2015). Laboratory experiments: professionals versus students. In: G. Fréchette and A. Schotter (eds), *Handbook of Experimental Economic Methodology*. Oxford: Oxford University Press.
- Friedman, J.W. (1971). A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1), 1–12.
- Georganas, S. and Nagel, R. (2011). Auctions with toeholds: An experimental study of company takeovers. *International Journal of Industrial Organization*, 29(1), 34–45.
- Goeree, J.K., Palmer, K., Holt, C.A., Shobe, W., and Burtraw, D. (2010). An experimental study of auctions versus grandfathering to assign pollution permits. *Journal of the European Economic Association*, 8(2–3), 514–525.
- Goette, L. and Schmutzler, A. (2009). Merger policy: what can we learn from experiments? In: J. Hinloopen and H.T. Normann, (eds), *Experiments and Competition Policy*. Cambridge, UK: Cambridge University Press.
- Goodwin, D. and Mestelman, S. (2010). A note comparing the capacity setting performance of the Kreps-Scheinkman duopoly model with the Cournot duopoly model in a laboratory setting. *International Journal of Industrial Organization*, 28(5), 522–525.
- Grubb, M. (2015). Behavioral consumers in industrial organization: An overview. *Review of Industrial Organization*, 47(3), 247–258.
- Gu, Y. and Wenzel, T. (2015). Putting on a tight leash and levelling playing field: An experiment in strategic obfuscation and consumer protection. *International Journal of Industrial Organization*, 42, 120–128.
- Guala, F. (2005). *The Methodology of Experimental Economics*. New York: Cambridge University Press.



472 *Handbook of game theory and industrial organization: applications*

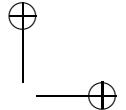
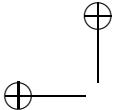
- Haan, M.A., Schoonbeek, L., and Winkel, B.M. (2009). Experimental results on collusion. In: J. Hinloopen and H.T. Normann (eds), *Experiments and Competition Policy*. Cambridge, UK: Cambridge University Press.
- Halbleib, D., Fehr, E., Goette, L., and Schmutzler, A. (2009). Self-reinforcing market dominance. *Games and Economic Behavior*, 67(2), 481–502.
- Hamilton, J.H. and Slutsky, S.M. (1990). Endogenous timing in duopoly games: Stackelberg or Cournot equilibria. *Games and Economic Behavior*, 2(1), 29–46.
- Hampton, K. and Sherstyuk, K. (2012). Demand shocks, capacity coordination, and industry performance: lessons from an economic laboratory. *The RAND Journal of Economics*, 43(1), 139–166.
- Harris, C. and Vickers, J. (1987). Racing with uncertainty. *Review of Economic Studies*, 54(1), 1–21.
- Henze, B., Schuett, F., and Sluijs, J.P. (2015). Transparency in markets for experience goods: experimental evidence. *Economic Inquiry*, 53(1), 640–659.
- Hoggatt, A.C. (1959). An experimental business game. *Behavioral Science*, 4(3), 192–203.
- Holmberg, P. (2007). Supply function equilibrium with asymmetric capacities and constant marginal costs. *The Energy Journal*, 28(2), 55–82.
- Holmberg, P. (2008). Unique supply function equilibrium with capacity constraints. *Energy Economics*, 30(1), 148–172.
- Holt, C.A. (1985). An experimental test of the consistent-conjectures hypothesis. *The American Economic Review*, 75(3), 314–325.
- Holt, C.A. (1995). Industrial organization: a survey of laboratory research. In: J. Kagel and A. Roth (eds), *Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- Holt, C. and Sherman, R. (1990). Advertising and product quality in posted-offer experiments. *Economic Inquiry*, 28(1), 39–56.
- Huck, S. (2009). Mergers in Stackelberg markets: an experimental study. In: J. Hinloopen (ed.), *Experiments and Competition Policy*. Cambridge, UK: Cambridge University Press.
- Huck, S. and Müller, W. (2000). Perfect versus imperfect observability – an experimental test of Bagwell's result. *Games and Economic Behavior*, 31(2), 174–190.
- Huck, S., Konrad, K.A., and Müller, W. (2001). Big fish eat small fish: on merger in Stackelberg markets. *Economics letters*, 73(2), 213–217.
- Huck, S., Konrad, K.A., Müller, W., and Normann, H.T. (2007). The merger paradox and why aspiration levels let it fail in the laboratory. *The Economic Journal*, 117(522), 1073–1095.
- Huck, S., Lünser, G.K., and Tyran, J.R. (2012). Competition fosters trust. *Games and Economic Behavior*, 76(1), 195–209.
- Huck, S., Lünser, G.K., and Tyran, J.R. (2016). Price competition and reputation in markets for experience goods: an experimental study. *The RAND Journal of Economics*, 47(1), 99–117.
- Huck, S., Müller, W., and Normann, H.T. (2001). Stackelberg beats Cournot – on collusion and efficiency in experimental markets. *The Economic Journal*, 111(474), 749–765.
- Huck, S., Müller, W., and Normann, H.T. (2002). To commit or not to commit: endogenous timing in experimental duopoly markets. *Games and Economic Behavior*, 38(2), 240–264.
- Huck, S., Müller, W., and Vriend, N.J. (2002). The East End, the West End and King's Cross: On clustering in the four-player Hotelling game. *Economic Inquiry*, 40(2), 231–240.
- Huck, S., Normann, H.T., and Oechssler, J. (2004). Two are few and four are many: number effects in experimental oligopolies. *Journal of Economic Behavior and Organization*, 53(4), 435–446.
- Isaac, R.M. and Reynolds, S.S. (1988). Appropriability and market structure in a stochastic invention model. *Quarterly Journal of Economics*, 103(4), 647–671.
- Jung, Y.J., Kagel, J.H., and Levin, D. (1994). On the existence of predatory pricing: An experimental study of reputation and entry deterrence in the chain-store game. *The RAND Journal of Economics*, 25, 72–93.
- Kalayci, K. (2015a). Price complexity and buyer confusion in markets. *Journal of Economic Behavior and Organization*, 111, 154–168.
- Kalayci, K. (2015b). Confusopoly: competition and obfuscation in markets. *Experimental Economics*, 19(2), 1–18.
- Kalayci, K. and Potters, J. (2011). Buyer confusion and market prices. *International Journal of Industrial Organization*, 29(1), 14–22.
- Kerschbamer, R., Sutter, M., and Dulleck, U. (2015). How social preferences shape incentives in (experimental) markets for credence goods. *The Economic Journal*, 127(600), 393–416.
- Klempner, P. and Meyer, M. (1989). Supply function equilibria in oligopoly under uncertainty. *Econometrica*, 57(6), 1243–1277.
- Kreps, D.M. and Scheinkman, J.A. (1983). Quantity precommitment and Bertrand competition yield Cournot outcomes. *The Bell Journal of Economics*, 14(2), 326–337.
- Kreps, D.M. and Wilson, R. (1982). Reputation and imperfect information. *Journal of Economic Theory*, 27(2), 253–279.
- Kruse, J.B., Rassenti, S., Reynolds, S.S., and Smith, V.L. (1994). Bertrand-Edgeworth competition in experimental markets. *Econometrica: Journal of the Econometric Society*, 62(2), 343–371.



- Kübler, D. and Müller, W. (2002). Simultaneous and sequential price competition in heterogeneous duopoly markets: experimental evidence. *International Journal of Industrial Organization*, 20(10), 1437–1460.
- Le Coq, C. and Orzen, H. (2006). Do forward markets enhance competition? Experimental evidence. *Journal of Economic Behavior and Organization*, 61(3), 415–431.
- Le Coq, C. and Sturluson, J.T. (2012). Does opponents' experience matter? Experimental evidence from a quantity precommitment game. *Journal of Economic Behavior and Organization*, 84(1), 265–277.
- Leufkens, K. and Peeters, R. (2011). Price dynamics and collusion under short-run price commitments. *International Journal of Industrial Organization*, 29(1), 134–153.
- Lindqvist, T. and Stenrek, J. (2005). The insiders' dilemma: an experiment on merger formation. *Experimental Economics*, 8(3), 267–284.
- Lynch, M., Miller, R.M., Plott, C.R., and Porter, R. (1986). Product quality, consumer information and "lemons" in experimental markets. In: P.M. Ippolito and D.T. Scheffman (eds), *Empirical Approaches to Consumer Protection Economics*. Washington, DC: Federal Trade Commission, 251–306.
- Mago, S.D. and Dechenaux, E. (2009). Price leadership and firm size asymmetry: an experimental analysis. *Experimental Economics*, 12(3), 289–317.
- Mago, S.D. and Pate, J.G. (2009). An experimental examination of competitor-based price matching guarantees. *Journal of Economic Behavior and Organization*, 70(1), 342–360.
- Mason, C.F. and Nowell, C. (1992). Entry, collusion, and capacity constraints. *Southern Economic Journal*, 58(4), 1002–1014.
- Mason, C.F. and Phillips, O.R. (1997). Information and cost asymmetry in experimental duopoly markets. *Review of Economics and Statistics*, 79(2), 290–299.
- Mason, C.F., Phillips, O.R., and Nowell, C. (1992). Duopoly behavior in asymmetric markets: An experimental evaluation. *The Review of Economics and Statistics*, 74(4), 662–670.
- McKelvey, R.D. and Palfrey, T.R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1), 6–38.
- Mestelman, S. and Welland, D. (1988). Advance production in posted offer markets. *Journal of Economic Behavior and Organization*, 8(2), 249–264.
- Mestelman, S. and Welland, D. (1991). Inventory carryover and the performance of alternative market institutions. *Southern Economic Journal*, 57, 1024–1042.
- Milgrom, P. and Roberts, J. (1982). Predation, reputation, and entry deterrence. *Journal of Economic Theory*, 27(2), 280–312.
- Miller, R.M. and Plott, C.R. (1985). Product quality signaling in experimental markets. *Econometrica: Journal of the Econometric Society*, 53(4), 837–872.
- Morgan, J., and Várdy, F. (2004). An experimental study of commitment in Stackelberg games with observation costs. *Games and Economic Behavior*, 49(2), 401–423.
- Morgan, J., Orzen, H., and Sefton, M. (2006a). An experimental study of price dispersion. *Games and Economic Behavior*, 54(1), 134–158.
- Morgan, J., Orzen, H., and Sefton, M. (2006b). A laboratory study of advertising and price competition. *European Economic Review*, 50(2), 323–347.
- Müller, W. (2006). Allowing for two production periods in the Cournot duopoly: experimental evidence. *Journal of Economic Behavior and Organization*, 60(1), 100–111.
- Müller, W. and Normann, H.T. (2014). Experimental economics in antitrust. In: R. Blair and D. Sokol (eds), *The Oxford Handbook of International Antitrust Economics, Vol. 1*. Oxford: Oxford University Press.
- Müller, W., Spiegel, Y., and Yehezkel, Y. (2009). Oligopoly limit-pricing in the lab. *Games and Economic Behavior*, 66(1), 373–393.
- Muren, A. (2000). Quantity precommitment in an experimental oligopoly market. *Journal of Economic Behavior and Organization*, 41(2), 147–157.
- Normann, H.T. and Wenzel, T. (2015). Shrouding add-on information: an experimental study. Paper at the Annual Conference on Economic Development – Theory and Policy, Münster, Germany.
- Offerman, T., Potters, J. and Sonnemans, J. (2002). Imitation and belief learning in an oligopoly experiment. *The Review of Economic Studies*, 69(4), 973–997.
- Orzen, H. (2008). Counterintuitive number effects in experimental oligopolies. *Experimental Economics*, 11(4), 390–401.
- Peeters, R. and Strobel, M. (2009). Pricing behavior in asymmetric markets with differentiated products. *International Journal of Industrial Organization*, 27(1), 24–32.
- Potters, J. and Suetens, S. (2009). Cooperation in experimental games of strategic complements and substitutes. *The Review of Economic Studies*, 76(3), 1125–1147.
- Potters, J. and Suetens, S. (2013). Oligopoly experiments in the current millennium. *Journal of Economic Surveys*, 27(3), 439–460.
- Puzzello, D. (2008). Tie-breaking rules and divisibility in experimental duopoly markets. *Journal of Economic Behavior and Organization*, 67(1), 164–179.

474 *Handbook of game theory and industrial organization: applications*

- Reynolds, S. (2000). Durable goods monopoly: laboratory market and bargaining experiments. *RAND Journal of Economics*, 31(2), 375–394.
- Sacco, D. and Schmutzler, A. (2011). Is there a U-shaped relation between competition and investment? *International Journal of Industrial Organization*, 29(1), 65–73.
- Saloner, G. (1987). Cournot duopoly with two production periods. *Journal of Economic Theory*, 42(1), 183–187.
- Santos-Pinto, L. (2008). Making sense of the experimental evidence on endogenous timing in duopoly markets. *Journal of Economic Behavior and Organization*, 68(3), 657–666.
- Sauermann, H. and Selten, R. (1959). Ein oligopolexperiment. *Zeitschrift für die gesamte Staatswissenschaft*, 115(3), 427–471.
- Selten, R. (1963). Ein Oligopolexperiment mit Preisvariation und Investition. In: H. Sauermann (ed.), *Beiträge zur experimentellen Wirtschaftsforschung*. Tübingen: J.C.B. Mohr, 103–135.
- Spiegler, R. (2011). *Bounded Rationality and Industrial Organization*. Oxford: Oxford University Press.
- Stigler, G.J. (1964). A theory of oligopoly. *The Journal of Political Economy*, 72(1), 44–61.
- Suetens, S. (2005). Cooperative and noncooperative R&D in experimental duopoly markets. *International Journal of Industrial Organization*, 23(1), 63–82.
- Suetens, S. (2006). R&D cooperation and strategic decision-making in oligopoly: an experimental economics approach. *Experimental Economics*, 9(2), 175–176.
- Van Damme, E. and Hurkens, S. (1997). Games with imperfectly observable commitment. *Games and Economic Behavior*, 21(1), 282–308.
- Várdy, F. (2004). The value of commitment in Stackelberg games with observation costs. *Games and Economic Behavior*, 49(2), 374–400.
- Yuan, H. and Krishna, A. (2011). Price-matching guarantees with endogenous search: a market experiment approach. *Journal of Retailing*, 87(2), 182–193.
- Zizzo, D.J. (2002). Racing with uncertainty: a patent race experiment. *International Journal of Industrial Organization*, 20(6), 877–902.



18. Empirical models of firms' R&D*

*Andrés Barge-Gil, Elena Huergo, Alberto López
and Lourdes Moreno*

1 INTRODUCTION

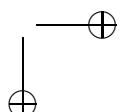
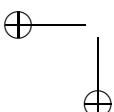
During the last few decades, economists have paid increasing attention to the role played by research and development (R&D) in modern economies. The origins of this research agenda are often credited to Solow's (1957) work on the importance of the total factor productivity growth as a measure of technological change, and to Nelson's (1959) and Arrow's (1962) works on the economics of knowledge creation (Encaoua et al., 2013). From a theoretical point of view, the tools of new game theory were applied to analyze the behavior and interactions of firms undertaking R&D. From an empirical point of view, some authors began the task of measuring and understanding the determinants and outcomes of R&D.

Nowadays, the empirical literature on R&D is abundant and still growing because of the increasing availability of micro-data. Taking this fact into account, the main purpose of this chapter is to provide a general view of three of the main topics covered by this literature: the determinants of firms' R&D investments, the link between R&D, innovation and productivity, and the studies trying to open up and examine the contents of the black box of R&D. To this aim, we build on already existing reviews, and complete them with some recent works. In addition, we pay special attention to some new lines of research not covered in previous reviews. For reasons of space, our analysis does not address other topics that appear to be worthwhile, like the relationship between R&D and employment, the outsourcing and/or offshoring of R&D, or R&D spillovers.

In the second section of the chapter, we review studies on the determinants of R&D, which are generally supported on Dorfman-Steiner-type (1954) models. In these models, profit-maximizing firms choose the level of R&D investment that equalizes the marginal revenue effect and the marginal cost effect of R&D expenditure. Regarding specific determinants of revenues and costs, most of the literature has focused on testing the so-called Schumpeterian hypotheses, which predict a positive relationship between size and R&D investment on the one hand, and between market power and R&D investment on the other. A related line of research also focuses on industry determinants of R&D: demand pull, technological opportunity and appropriability.

Much attention has recently been devoted to the role of public funding as a determinant of business R&D investment. There is considerable agreement that R&D activities suffer from market failures. First of all, the presence of information asymmetries and moral hazard

* This research has been partially funded by the Spanish Ministry of Economy and Competitiveness (project ECO2014-52051-R) and by the Autonomous Region of Madrid through project S2015HUM-3417 (INNCOMCON-CM), co-funded by the European Social Fund (European Union).



increases the cost of financing R&D relative to other investments. Second, the main output of R&D activities is knowledge, which shows some “public good” characteristics that make its full appropriation difficult. As a consequence, the amount of public funding for R&D is of considerable magnitude in most countries, and many studies have aimed to evaluate its effect. Despite the profusion of empirical literature about this subject (see David, Hall and Toole, 2000; Becker, 2013; and Zúñiga-Vicente et al., 2014, for a review), only a few papers propose structural models describing firms’ decisions as the analytical framework for their estimations. We pay special attention to these papers in our assessment.

In the third section of the chapter, we deal with studies analyzing the relationship between R&D, innovation and productivity. In this regard, we take into account three main approaches: the knowledge capital model, the model proposed by Crépon, Duguet and Mairesse (1998) and subsequently known as the CDM model, and a group of recent structural models.

In most early studies, an augmented production function with R&D capital (or R&D expenditures) is used to estimate the returns to R&D at the firm level. Later, this approach was improved by using a more complex modelization in which a demand equation is also included to control the bias generated in the estimation of a firm-level production function where output is proxied by deflated sales. In this case, the firm’s knowledge capital also affects the demand by improving product quality. Hall, Mairesse and Mohnen (2010) review the literature on returns to R&D based on different specifications of the production function at the plant, firm, industry and country levels since the 1960s. We summarize their review and include some new papers.

A second approach refers to the CDM model, which takes into account the fact that it is not innovation inputs but innovation outputs that increase productivity. Specifically, under this approach, productivity is explained by technological outputs and the latter by technological effort. Revisions of this approach have already been provided by Hall (2011) and Mohnen and Hall (2013). However, most papers included in these revisions use cross-sectional data and do not take into account dynamic linkages between innovation and economic performance or persistence in R&D activities. The availability of longer time series and new econometric methods has generated some recent empirical evidence that considers the timing of innovation and other dynamics aspects. There are different explanations for persistent behavior in R&D activities: sunk costs associated with the performance of R&D activities, the “success breeds success” hypothesis and the existence of dynamic increasing returns, among others. We summarize this empirical evidence, which has not been reviewed before.

Finally, we revise the results obtained in recent papers that develop structural models to relate productivity and R&D. In these studies, productivity is assumed to be unobservable and is modeled as a Markov process that depends on R&D expenditure or other endogenous firms’ decisions. In addition, to deal with the simultaneity problem, instead of estimating a production function, these authors model and estimate the firm’s dynamic decision to invest in R&D. Specifically, they propose a dynamic structural model of R&D demand where the expected benefit to the firm’s investment in R&D is inferred from the rational decision of R&D investment.

The fourth section of this chapter concerns itself with studies aimed at opening up the black box of R&D. In particular, we examine two lines of research. The first one distinguishes between the components of R&D, while the second analyzes the complementarities among different types of R&D activities.

One common drawback of much of the literature is that R&D is considered a single activity. However, the black box of R&D is composed of basic research, applied research and development activities. These activities are different in purposes, knowledge bases, the people involved and management styles, so their determinants and outcomes may be quite different. While some seminal studies were carried out on this topic in the 1980s, it has gained renewed attention in the last decade. To our knowledge, no systematic empirical review addresses this subject.

A related topic is that internal R&D is not carried out in isolation. On the contrary, in-house R&D activities should be understood as part of a more general strategy of the firm. In this regard, one important topic of the recent literature has focused on the analysis of the complementarities of internal R&D with other firm decisions related to technological activities. From an empirical point of view, the analysis of complementarities involves some specific challenges. We briefly summarize both the methodologies developed to address these challenges and the main empirical studies on complementarity in R&D activities.

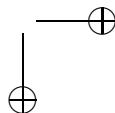
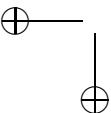
Finally, in Section 5, we provide the main conclusions of our analysis and suggest some avenues for further research.

2 R&D DETERMINANTS

The study of research and development (R&D) determinants is a classic topic in industrial organization. From a theoretical point of view, these studies are generally supported on Dorfman-Steiner-type (1954) models, in which profit-maximizing firms choose the level of R&D investment that equalizes the marginal revenue effect and the marginal cost effect of R&D expenditure (Kamien and Schwartz, 1970, 1976, 1978, 1982; Needham, 1975; Scherer, 1980).¹ The specific relation among the key variables depends on whether the R&D investment is considered a demand-increasing strategy and/or a cost-reducing strategy. In particular, Cohen and Levin (1989) emphasize that the impact of price elasticity of demand will be ambiguous in empirical studies that do not distinguish between product and process innovation: the gains from process innovation, which is associated with cost-reducing strategies, will be larger the more elastic demand is (Kamien and Schwartz, 1970), while the gains from product innovation, which is related to demand-increasing strategies, will be larger the more inelastic demand is, given that inelastic demand could amplify the gains from a rightward shift in the demand curve (Spence, 1975).

With the emergence of new firm-level and project-level databases at the end of the twentieth century, the study of the determinants of R&D investment faced a revival. As in previous works, most of the authors tended to distinguish between demand-side effects and cost-side effects of R&D. The implicit framework in these analyses assumes that, for each planning period and R&D project, the level of R&D expenditure is the result of equalizing the marginal

¹ For instance, following Needham (1975), the profit-maximizing R&D expenditure condition can be written as follows: $\frac{R}{P\gamma} = \frac{\varepsilon_R + \varepsilon_{conj} \cdot \varepsilon_{Rr}}{\varepsilon_d}$, where ε_d is the price elasticity of demand for the firm's product, ε_R is the elasticity of the quantity demanded of the firm's product with respect to the firm's R&D expenditure, ε_{conj} is the elasticity of the rival's R&D expenditure with respect to the firm's own R&D expenditure and ε_{Rr} is the elasticity of demand for the firm's product with respect to the rival's R&D expenditure.



cost of R&D, MCR , and the marginal revenue of R&D, MRR . Following David et al. (2000), we can represent this setting through the equations:²

$$MRR = f(R, \mathbf{X}) \quad (18.1)$$

$$MCR = g(R, \mathbf{Z}) \quad (18.2)$$

where R is the level of R&D expenditures. In this simplified model \mathbf{X} stands for the vector of other variables that can affect the distribution of project rates of return, like technological opportunities and appropriability conditions, while \mathbf{Z} represents the vector of other variables that determine the marginal cost of R&D, such as those related to access to bank financing or venture capital, to macroeconomic conditions or to technological public policy.³

Where MRR and MCR are equal, we find the firm's profit-maximizing level of R&D investment, R^* :

$$R^* = h(\mathbf{X}, \mathbf{Z}) \quad (18.3)$$

Departing from this kind of model, the use of a micro-level perspective in empirical analysis allows the testing of new hypotheses. One of the most relevant during the last few decades is the complementarity or substitutability relation between public R&D and private R&D. This concern about the effect of public aid on business R&D points out the relevance of taking into account not only the determinants of R&D-intensive margins, but also the determinants of extensive margins, that is, the propensity to perform R&D (González, Jaumandreu and Pazó, 2005; Takalo, Tanayama and Toivanen, 2013a; Arqué-Castells, 2013; Arqué-Castells and Mohnen, 2015). The evidence obtained about extensive margins supports the existence of sunk costs of R&D that act as a barrier to entering R&D markets, especially for small firms. This would be one of the explanations for R&D persistence, and justifies the interest in dynamic approaches to model R&D decisions. Some of these topics will be analyzed in more detail in the following sections.

There are several estimation issues involved in the estimation of the R&D equation: unobservable heterogeneity, dynamics and persistence, endogeneity and parameter heterogeneity. First, a number of unobservable characteristics, such as managerial ability or culture, usually exist. This issue is dealt with by using some kind of transformation (within or differencing) that wipes out individual (time-invariant) effects.

Second, R&D is an investment that should be analyzed in a dynamic framework because it shows high adjustment costs. Accordingly, firms tend to smooth their R&D investment over time (Lach and Schankerman, 1989). Most R&D studies use standard investment equation methodology to incorporate adjustment cost dynamics into the static R&D model. The main approach is a neoclassical accelerator model with ad hoc dynamics (Mairesse, Hall and Mulkay, 1999), introducing a lagged dependent variable. This model is sometimes expressed in error correction form in order to take explicit account of short-term versus long-term effects (Bond, Harhoff and Van Reenen, 2005). Recent developments distinguish adjustment costs of

² Throughout the chapter we slightly change the original notation in reference papers to keep the notation as homogeneous as possible in all sections.

³ Prior antecedents to this model can be found in Anderson (1967) and Howe and McFetridge (1976). See also a reference to this model in Martin (2010).

those firms starting R&D from adjustment costs of those firms already investing in R&D (Peters et al., 2013).⁴

Third, R&D endogeneity may take place for a number of reasons: (i) it may be correlated with unobserved time-variant effects; (ii) it may respond to expectations regarding future technological shocks; (iii) the random shocks to current R&D (ε_{it}) may affect the future values of the explanatory variables; or, (iv) it may be determined simultaneously with other firm characteristics included in X_{it} . These problems have usually been addressed using IV-GMM (generalized method of moments), where the model is estimated in first differences and the lagged levels are used as instruments, and system-GMM where the equation in levels is added and the lagged differences are used as instruments (Bond et al., 2005).

Fourth, the models usually assume that the effect of the different determinants is homogeneous across firms. If one believes they are heterogeneous, then the estimated coefficients reflect average effects within the sample. Some authors have explored whether the effect of R&D determinants may differ across different types of firms (e.g., small vs large, low-tech vs high-tech or young vs mature).

The rest of this section is organized as follows. First, we review the literature on the classical determinants of R&D. Second, we summarize the more recent structural studies focusing on the role played by public funding.

2.1 Classical Determinants

The main classical determinants of R&D investment are firm size, market power and industry determinants, such as appropriability, demand pull and technological opportunity. The role played by size and market power in firm investment has been analyzed since Schumpeter (1939, 1942) while the focus on industry determinants was raised by a series of papers in the 1980s (Levin, Cohen and Mowery, 1985; Cohen, Levin and Mowery, 1987; Cohen and Levin, 1989).

2.1.1 Size

There are several arguments that support a positive relationship between firm size and R&D investment. The main one is related to the availability of R&D funding. Larger firms are usually able to generate larger cash flows to internally finance R&D investment, and they have better access to imperfect capital markets. In addition, R&D investment usually shows scale and scope economies.

The empirical analysis of the size–R&D relationship has been a matter of interest for decades, leading to the development of some stylized facts (Cohen and Klepper, 1996; Cohen, 2010): (i) the likelihood of performing R&D increases with firm size; (ii) R&D rises monotonically, and typically proportionally, with firm size (within industries and among R&D performers); and (iii) small firms account for a larger share of innovations and patents than expected according to their R&D investment.

More recent studies focus on directly analyzing the relationship between R&D and the motives behind the importance of size. Many studies focus on indicators of internal and external availability of funding, usually concluding that availability of funding is positively related to R&D investment effect (Bloch, 2005; Brown et al., 2009; Czarnitzki and Hottenrott,

⁴ This issue will be addressed in detail in Section 3.4.

2011; Borisova and Brown, 2013).⁵ Some evidence has been developed that this relationship is stronger in the USA and the UK than in continental Europe (Hall, 2002; Bond, Fazzari and Peterson, 2005), because of a different structure of capital markets and different corporate attitudes towards uncertainty. In turn, Henderson and Cockburn (1996) find strong evidence of the existence of scale and scope economies using program-level data from the pharmaceutical industry.

2.1.2 Market power

The relationship between market competition and R&D investment is less straightforward. Theory postulates two different effects. On the one hand, a decrease in market power may reduce the incentive to invest in R&D because the firm would be less able to extract the rent resulting from innovation output (Grossman and Helpman, 1991; Aghion and Howitt, 1992). On the other hand, innovation could partially displace oligopolistic rents, thus reducing R&D incentives for firms with high market power (Arrow, 1962), or it may be used as a strategic variable to face increased competition (Spencer and Brander, 1983). The theoretical controversy has stimulated empirical research. Some authors have obtained a positive influence of market power on innovation (Crépon et al., 1998; Blundell, Griffith and Van Reenen, 1999), while other authors have found a negative one (Geroski, 1990; Harris, Rogers and Siouclis, 2003). As a third possibility, Aghion et al. (2005) find that the relationship between product market competition and innovation is an inverted U-shape. All in all, Cohen (2010) concludes that what emerges from previous empirical evidence is that market power does not seem to play an important, independent role in affecting R&D.⁶

2.1.3 Industry determinants

Since the late 1980s some studies (Levin et al., 1985; Cohen et al., 1987; Cohen and Levin, 1989; Geroski, 1990) have placed great attention on industry-specific determinants of R&D, such as appropriability, demand and technology opportunity.

Appropriability issues have received most of the attention. Firms should be able to appropriate returns sufficient to make their investment worthwhile (Levin et al., 1985). Appropriability may take place using formal or informal methods. Although most of the studies have focused on the role played by the patenting system, empirical evidence suggests that lead time, secrecy and complementary assets are the most employed methods to appropriate innovation results, especially outside some specific industries, such as pharmaceuticals or medical equipment (Hall et al., 2014). However, the idea that more appropriability is related to more innovation effort has been questioned (Bessen and Maskin, 2009; Lerner, 2009). The reason is that knowledge spillovers and own R&D may be complements. That is, higher appropriability means a decrease of spillovers, which may lead to a net decrease of own R&D (Hall et al., 2014). Empirical evidence is not conclusive. On the one hand, firms in industries with high formal appropriability are found to invest more in R&D (Hall and Sena, 2014). On the other, the increase in formal property rights has been found to increase only patents but not

⁵ Some studies have not found a significant effect (Bond et al., 2005).

⁶ In a recent study, Beneito et al. (2015) argue that these contradictory results are driven by the impossibility of finding an accurate measure of market power. Instead, they use indicators of the fundamental determinants of competitive pressure (product substitutability, size of the market and ease of entry) and distinguish between product and process innovation. They conclude that greater product substitutability and higher costs of entry induce greater process innovation and lower product innovation, while market enlargement spurs both types of innovation.

R&D in the semiconductor industry (Hall and Ziedonis, 2001), while a strengthening of legal protection of trade secrets has been found to reduce R&D in manufacturing firms (Png, 2017).

Some studies qualify these results. Arora and Ceccagnoli (2006) show that stronger patent protection helps firms lacking complementary assets to appropriate from innovation output through licensing. Arora, Ceccagnoli and Cohen (2008) find that patenting is a net cost for the typical invention but very valuable for a subset of inventions and, consequently does provide an incentive for R&D. Czarnitzki and Toole (2011) find that patents increase R&D in German firms through a reduction in market uncertainty, and Duguet and Lelarge (2012) find that patents clearly promote R&D and product innovations but not process innovations in French firms.

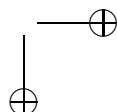
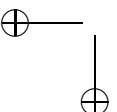
Regarding the role of demand and technology opportunity, they are the subjects of an old debate: the seminal works by Schmookler (1962) pointed out the critical importance of demand as a driver of innovation. The underlying assumption is that a common pool of knowledge and capabilities is available to all industries, and therefore large and growing markets provide higher incentives to invest in innovation as these markets offer higher returns for the investment (Cohen and Levin, 1989). On the other hand, some authors supported the view that technological opportunities were the driving force of technological change. Technological opportunities comprise the set of possibilities for advancing the knowledge frontier and may be measured in terms of the distribution of values of improved production-function or product-attribute parameters that may be attained through R&D or, alternatively, as the distribution of returns to R&D, given demand conditions and the appropriability regime (Klevorick et al., 1995). That is, at prevailing input prices, innovation is “easier” (less costly) in some industries than in others (Cohen and Levin, 1989), so more R&D is found in these industries. Technological opportunities are usually considered exogenous to a firm’s decisions. Some authors, however, argue that its influence over technological input is not clear. The reason is that technological opportunity may raise the average product of R&D without raising its marginal product, and therefore may not increase R&D investment (Klevorick et al., 1995).

Empirical studies support the importance of demand pull (Cohen et al., 1987; Acemoglu and Linn, 2004), although this support is not as strong as Schmookler thought (Geroski and Walters, 1995). On the other hand, empirical studies have usually agreed that technological opportunities are very influential in driving technological change (Raymond et al., 2010; Graevenitz, Wagner and Harhoff, 2013). Accordingly, the old debate has been solved, concluding that both are important determinants of R&D investments, although the fraction of variance explained by technological opportunity seems larger.

2.1.4 The Role of Public Funding

The main justification for public support of R&D activities is the correction of market failures. Because of the presence of information asymmetries and moral hazard, innovating firms usually face a higher cost of R&D finance with respect to ordinary investment and show a lower level of private external financing (Hall, 2002; Hall and Lerner, 2010). In addition, the “public good” nature of knowledge prevents full appropriation, which pushes private R&D investment below the socially optimal level.

Taking this into account, there is abundant literature analyzing whether public R&D spending complements or displaces private R&D spending. When public support consists of subsidies or loans, testing for complementarity relies on determining whether public



R&D induces additional private R&D investment beyond the level that would have been performed anyway.

The spectrum of empirical methodologies for performing this analysis is wide, including specific techniques to control for potential endogeneity of public support, firm heterogeneity or non-linearities, among other issues. Interested readers are referred to the reviews by David et al. (2000), Becker (2013) and Zúñiga-Vicente et al. (2014). As these reviews point out, the evidence is ambiguous, with results supporting both a (total or partial) crowding-out and a crowding-in (additionality) effect of public subsidies on private R&D investments.

Despite this wealth of empirical literature about the impact of public R&D funding on business R&D, only a few papers propose structural models describing firms' decisions as the analytical framework for their estimations. Outstanding exceptions are the works by González et al. (2005), Takalo et al. (2013a, 2013b) and Arqué-Castells and Mohnen (2015).

González et al. (2005) model firms' decisions about R&D extensive and intensive margins when some government support can be expected. In their model, each firm is a product-differentiated competitor capable of shifting the demand for its product by enhancing product quality through R&D. To decide the level of R&D expenditures, the firm maximizes expected profits, $E[N(R) - (1-s)^\beta R]$, where firm net revenue, N , is a function of R&D expenditures, R , with s being the subsidized fraction of these R&D expenditures. In the equilibrium, the optimal non-zero R&D effort can be expressed as follows:

$$e^* = \frac{R^*}{P^* Y^*} = \frac{R \partial Y}{Y \partial R} / \left(-\frac{P \partial Y}{Y \partial P} E[(1-s)^\beta] \right), \quad (18.4)$$

where E denotes the expectation over s values and β stands for the level of expenditure efficiency of public funds.

This Dorfman-Steiner-type (1954) expression reflects both the optimal effort of performing and non-performing firms. It will only be observed if it surpasses a threshold effort \bar{e} that corresponds to the level of expenditure that makes the firm indifferent between performing R&D or not. Below this threshold, R&D costs are not completely recovered by means of the sales increment. Notice that, as expected subsidies reduce the cost of R&D, they affect both the decision to undertake innovative activities and the size of planned R&D expenditures of performing firms.

As González et al. (2005) explain, this framework naturally leads to a Tobit-type modeling of a censored variable for estimating the model parameters and, particularly, the effect of subsidies. The econometric model consists of the following equations:

$$e^* = -\beta \ln(1 - s^e) + z_1 \beta_1 + u_1 \quad (18.5)$$

$$\bar{e} = z \beta_2 + u_2, \quad (18.6)$$

where e^* is only observed when $e^* - \bar{e} > 0$. z (which contains at least all variables in z_1), stands for the vector of variables that determine the spending profitability threshold, that is, demand characteristics, technological opportunities and set-up costs of R&D projects. The error term u_1 is assumed to be autocorrelated, while the error term u_2 is supposed to be independent and identically distributed over time. S^e is the expectation for s , which is unobservable and must be previously estimated. To do so, the expectation is decomposed in

the product of the conditional expectation of receiving a grant, $P(s > 0|z_p)$, and the expected value of the subsidy conditional on z_p and its granting, $E(s|z_p, s > 0)$. These two components are estimated using, respectively, a probit and an ordinary least squares (OLS) specification.

Given that subsidies are presumably granted by agencies according to the effort and performance of firms, for the estimation of this model González et al. (2005) apply methods for dealing with selectivity and endogeneity in a context that allows for autocorrelated errors. One of the main conclusions of their analysis is that subsidies can stimulate non-R&D-performing firms to start investing in R&D.

Takalo et al. (2013b) complement the framework of González et al. (2005) with the structural modeling of the subsidy-application decision of firms, each one with a unique R&D project, and the subsidy-granting decision of the public agency. In particular, the subsidy program is modeled as a four-stage game of incomplete information between both players. In stage 0, the players' types (denoted by ε and η , respectively, for the project/firm and the agency) are determined.⁷ In stage 1, the firm decides whether or not to apply for a subsidy. In stage 2, the agency grades the proposal and learns its type. Finally, in stage 3, the firm chooses the R&D investment with or without the subsidy. As in González et al. (2005), the optimal R&D investment (intensive margin) is obtained through the first-order condition of the firm's profit-maximizing problem.

As Takalo et al. (2013b) point out, given that the goal of their approach is to derive equations that could be empirically estimated, they "model the players' payoffs by more specific functional forms that would be necessary from a purely theoretical point of view" (p. 257). In particular, the objective function of the firm and the agency's expected utility from an applicant's project are expressed, respectively, as:

$$\Pi(R(s), s, X, \varepsilon) = \exp(X\beta + \varepsilon) \ln R(s) - (1 - s)R(s) \quad (18.7)$$

$$U(R(s), s, X, Z, \varepsilon, \eta) = V(R(s), Z, \eta) + \Pi(R(s), s, X, \varepsilon) - gsR(s) = F, \quad (18.8)$$

where S stands for the subsidy rate. In equation (18.8), the first term, $V(\cdot)$, captures the (domestic) spillovers of the project beyond the firm's profits, the direct costs of the subsidy ($gsR(s)$) and the fixed costs of applying and processing the application (F). g denotes the constant opportunity cost of agency resources, while X and Z stand for vectors of observable firm characteristics.⁸

Under this framework, Takalo et al. (2013b) prove that there is a unique perfect Bayesian equilibrium of the game. The econometric implementation of this model relies on the estimation of four types of equations regarding the firms' application and R&D investment decisions, the agency's subsidy rate decision and the grading process. The estimation of this multi-equational system with project-level data from Finland allows them to quantify the benefits and costs of the R&D subsidy program. They report four main findings: (i) the expected effects of subsidies are very heterogeneous; (ii) estimated application costs vary greatly, and shocks to application costs and marginal profitability of R&D are positively correlated; (iii) spillover effects of subsidies are somewhat smaller than effects on firm profits; and (iv) the expected rate of return on the subsidy program is about 30 to 50 percent.

⁷ The type of project and the agency's type corresponding to each project are drawn from common knowledge (joint) distributions and constitute the unobservables of the econometric model.

⁸ See the complete set of equations in the original paper.

Two strong assumptions in the depicted model are the lack of fixed R&D costs and the absence of credit rationing. Takalo et al. (2013a) specifically extend this model by including a cost of external funding and fixed costs of R&D projects. To overcome liquidity problems and given that public subsidies are paid ex post, these authors assume that the firm can raise funding from financial markets, which are competitive with free entry of identical financiers and an unlimited supply of funds. Therefore, the contract between the firm and the financier can be defined as the one that maximizes the firm's payoff subject to a financier's zero profit condition. In this framework, the presence of fixed costs of R&D allows us to determine the effects of R&D subsidies at the extensive margin, where the firms decide whether or not to invest in R&D. In addition, the cost of external finance reduces the optimal subsidy rate at the intensive margin, while the association is the opposite at the extensive margin. Takalo et al. (2013a) also derive necessary and sufficient conditions for the existence of additionality effects of public support to private R&D. In particular, they show that additionality at the intensive margin inversely depends on the spillover rate. As a consequence, the relationship between additionality and welfare may be ambiguous.

Focusing explicitly on the firm's optimal R&D decisions, Arqué-Castells and Mohnen (2015) consider an additional dimension: the existence of sunk entry costs of R&D. The firm would incur these costs only the first time that it engages in R&D. Classical subsidy granting schemes consist of subsidies seeking to increase the intensive margin, i.e., to promote the R&D effort of regular R&D performers. The existence of sunk costs also provides a justification for the use of "extensive" subsidies, i.e., subsidies seeking to expand the base of R&D performers through the coverage of these costs.

The model framed by Arqué-Castells and Mohnen is similar to the ones in González et al. (2005) and Takalo et al. (2013a, 2013b). However, Arqué-Castells and Mohnen consider a dynamic model in which firms decide whether to start, continue or stop performing R&D depending on R&D "intensive" or "extensive" subsidies. Through this model, they characterize the firm's optimal participation strategy in terms of two subsidy thresholds, which determine R&D entry and continuation. Using firm-level data on Spanish manufacturing firms, Arqué-Castells and Mohnen (2015) are able to compute these thresholds through the estimation of a dynamic panel data Type II Tobit model with two equations: the R&D participation equation and the R&D investment equation. Simulation on this estimated model suggests that one-shot trigger subsidies positively affect the share of R&D performers and average R&D expenditures. In addition, extensive subsidies might induce permanent effects, as R&D performers in a given period are more likely than non-performers to undertake R&D activities in the next period.

3 R&D INNOVATION AND PRODUCTIVITY

The analysis of productivity growth and its determinants is a classic topic in industrial economics. There are a large number of papers that study this question from an empirical point of view, pointing out the performance of technological activities as an essential source of firms' growth. Specifically, many authors have analyzed the relationship between R&D activities and productivity, finding, in general, a positive and significant effect of R&D on productivity, although with different magnitudes depending on the methodology employed and the level of analysis. In this respect, as Mairesse and Sassenou (1991) point out, the issue

is not so much the question of whether or not such a relationship exists, but whether or not econometric studies can characterize such a relationship in a satisfactory and useful manner.

R&D activity can affect productivity in different ways. It can reduce the production costs or increase the quality of the goods that exist in the markets. But it can also generate new goods. These effects can cause price reductions, margin increases and reallocations, not only in terms of factors, but also in terms of firm entry and exit in the market. In this sense, it is difficult to accurately measure the impact of R&D because it affects both supply and demand of products. In addition, R&D undertaken by other firms in the own sector (or different sectors and countries) can generate positive spillovers in other firms of the same sector (or different sectors or different countries) and should also be taken account. In our revision, we are going to distinguish between three main approaches: the knowledge capital model, the CDM model⁹ and recent structural models.

Most early studies follow Griliches (1979), using an augmented production function with R&D capital (or R&D expenditure) to estimate the returns to R&D at the firm level. They focus on the supply side, estimating a Cobb-Douglas production¹⁰ in levels or in log differences, which allows them to compute the impact of R&D capital on the level or the growth of the productivity. In this context, under the assumptions of constant returns to scale and perfect competition, some papers use the growth accounting framework, which allows relating the growth of total factor productivity (TFP) to R&D in a non-parametric way. This framework will be known as the knowledge capital model.

Later, following Hall (1988), Klette (1996) proposes a combination of non-parametric and parametric productivity analysis to take into account not only the existence of mark-ups as in Hall (1988) but also the treatment of scale economies. In addition, and as in Klette and Griliches (1996), he tries to control the bias generated in the estimation of firm-level production functions when output is proxied by deflated sales, based on a common deflator across firms. In this case, biases occur in situations where firms compete in an imperfectly competitive environment in which prices will reflect idiosyncratic differences in cost and in market power across firms. Some of these differences are generated by the innovation in creating and/or increasing that market power. To control this, they add a model of product demand to the model of producer behavior. The firm's knowledge capital affects the demand through improved product quality. They rewrite the production function in terms of revenue rather than real output assuming an isoelastic demand function and combining the production function with the demand equation.

In the previous approach, the variable that measures the technological activity and enters in both supply and demand equations is R&D capital (or R&D expenditures), which is an input measure. Recently, the idea that the growth of firms is more related to the results of technological activities than to the inputs used in them has generated some studies that directly analyze the impact of technological outputs (process, product or organizational innovations, patents, etc.) on firms' productivity. Specifically, Crépon et al. (1998) developed a multi-equational model that explains productivity by technological outputs and the latter by technological effort (innovation input). Since the appearance of this seminal paper, many researchers have applied the same methodology to different European countries, using

⁹ The model proposed by Crépon et al. (1998) and subsequently known as the CDM model.

¹⁰ A smaller number of papers estimate a system of factor demand equations derived from a dual (cost function) representation of technology.

essentially cross-sectional data from Community Innovation Surveys (CIS data). Most of them estimate the equations of this model in a recursive way, although some recent papers estimate the equations simultaneously.

More recently, some authors developed structural models to relate productivity and R&D. Specifically, they assume that productivity is unobservable and model it as a Markov process that depends on the R&D expenditure or other endogenous decisions of the firms (Aw, Roberts and Xu, 2011; Doraszelski and Jaumandreu, 2013). In this context, Aw et al. (2011), Roberts and Vuong (2013) and Peters et al. (2013) propose a dynamic structural model of R&D demand where the expected benefit to the firm's investment in R&D is inferred from the rational decision of R&D investment.

In Sections 3.1 and 3.2, the general theoretical framework of the classical approach is summarized as the empirical results derived from it. In Sections 3.3 and 3.4, the CDM model and the new papers taking into account the persistence inside the structure of this model are depicted. Finally, Section 3.5 presents some recent structural models that relate R&D and productivity.

3.1 The Knowledge Capital Model

Griliches (1979) introduced a basic framework where the production function is augmented with a knowledge capital term. In this case, the productivity equation starts by assuming a production function for firm i in year t of the type:

$$Y_{it} = A_{it}F(L_{it}, K_{it}, M_{it}, G_{it}), \quad (18.9)$$

where Y denotes the quantity of output, L , K and M are, respectively, the quantities of labor, physical capital and materials, G denotes the quantity of knowledge capital, and A represents the level of efficiency reached by the firm (or a productivity shifter). Griliches (1979) proposed computing the knowledge capital from the accumulation of R&D expenditures over time, i.e., to construct a variable that measures the stock of R&D capital owned by a firm. Specifically, he used the perpetual inventory method, and this method remains the most widely used.

Assuming a Cobb-Douglas production function for equation (18.9) and taking logarithms, we obtain two alternative linear regressions (in levels and first differences) to be estimated:

$$y_{it} = a_{it} + \alpha_l l_{it} + \alpha_k k_{it} + \alpha_m m_{it} + \gamma g_{it} + u_{it} \quad (18.10)$$

$$\Delta y_{it} = \Delta a_{it} + \alpha_l \Delta l_{it} + \alpha_k \Delta k_{it} + \alpha_m \Delta m_{it} + \gamma \Delta g_{it} + \Delta u_{it}, \quad (18.11)$$

where lower-case letters denote logarithms of the variables and Δ denotes first differences of the variables. The parameters α_l , α_k and α_m measure the output elasticities with respect to the traditional inputs, and γ is the output elasticity with respect to the knowledge capital of the firm. In this context, γ is the main parameter of interest and it may reflect the impact of process innovation. In this sense, the knowledge capital is expected to have a direct positive effect on productivity because new processes reduce production costs. In addition, process innovations can also have indirect effects: if cost reductions are translated to prices, a big enough increase in sales to compensate price decreases can generate additional productivity improvements in the presence of returns to scale. However, the translation to the price depends on the competition in the market.

Estimation of firm-level production functions (such as equations (18.10) and (18.11)) involves a number of econometric issues, including simultaneity (endogeneity of the firm's input choices) and selection bias (endogenous exit).¹¹ Apart from these econometric issues, the estimation of production functions is full of challenges, especially related to data problems and, in particular, to measurement error in inputs and output. In the context of the knowledge capital model, the construction of R&D capital stock is problematic. As we said before, this capital is usually computed using the perpetual inventory method, but this method has major drawbacks in practice.¹² First, the depreciation rate is unknown and usually considered to be constant over time. Second, there is a problem related to the initial conditions for R&D capital stock.

Other measurement issues are related to quality changes in both inputs and output. If the prices used to deflate nominal inputs do not include quality adjustment, the real input would be undervalued and the TFP would be overestimated. This problem is clear in the case of ICT equipment but can also be applied to labor input, whose productivity can be increased over time. In this regard, if inputs are quality adjusted, they can capture the effects of innovation on them. With respect to the output, if the price doesn't incorporate the quality changes, the real output is undervalued when an industry output deflator is used. Instead, the quality improvement is reflected as an increase in the nominal revenue.

For the output measure, there is an additional problem related to non-competitive pricing. Market power across firms can be different and the differences are associated with the innovator behavior of firms: product innovations give a market-power position to the firms that allow them to sell at prices higher than competitive prices.¹³ As in the case of quality improvement, if there are no firm prices, the use of industry output deflators implies that part of the estimated productivity reflects price effects.

A second way to calculate TFP growth is the accounting growth approach, which assumes some assumptions that allow calculating the elasticities from observables. Specifically, under assumptions of constant returns to scale with respect to L , K and M and competitive markets, the input elasticities are the shares of revenue received by each of the factors. Even in the context of imperfect markets, cost minimization implies that input elasticities equal cost shares, s_i . Therefore, the Solow residual can be rewritten as:

$$\tilde{\theta}_{it} = y_{it} - (s_l \cdot l_{it} + s_k \cdot k_{it} + s_m \cdot k_{it}) \quad (18.12)$$

The main problems of this methodology are the non-fulfillment of the previous assumptions associated with this kind of model (constant returns to scale, instant adjustment of the inputs, competitive markets) and the interpretation of TFP. Note that it picks up everything not captured by the labor productivity, capital intensity, materials as changes in the firms' efficiency, capacity utilization or measurement errors of the variable (output and inputs). Nevertheless, some papers use this approach to estimate TFP, including some variables reflecting the non-fulfillment of the assumed assumptions, x_{it} , along with other control

¹¹ Van Beveren (2012) provides a detailed overview of both the econometric issues that arise when estimating total factor productivity in a production function framework at the firm level, and the existing (parametric and semi-parametric) techniques designed to overcome them.

¹² See Van Beveren (2012) for a detailed discussion.

¹³ In Dobbelaere and Mairesse (2010, 2013), they also consider monopsonistic competition, which allows firms to hire some of their inputs below competitive prices.

variables like the cycle. In this context, to estimate the impact of knowledge capital on TFP, equation (18.11) can be replaced by:¹⁴

$$\tilde{\theta}_{it} = \lambda \cdot g_{it} + x_{it}' \beta + v_{it} \quad (18.13)$$

Hall et al. (2010) review the literature on returns to R&D based on different specifications of the production function previously revised at the plant, firm, industry, or country levels since the 1960s. Most papers refer to pooled (or temporal) estimates on firm data that use the level production function, although they also consider estimates from growth rates regressed on R&D intensity. When the production function is estimated in first-differenced form, they find a substantial downward bias to the R&D coefficient. This bias can be mitigated by imposing constant returns. They find that research elasticities range from 0.01 to 0.25. The cross-sectional estimates are higher than the panel data estimates, which in some cases are statistically non-significant. One explanation might be that measurement errors have a much more serious impact on growth rates than on the levels of variables. The rate of return is obtained by multiplying the estimated elasticity by the average output–R&D capital ratio in the sample. The R&D rates of return in developed economies during the past half-century have been strongly positive and are more likely to be in the 20–30 percent range.

Wieser (2005) surveys the empirical literature on firm-level R&D and productivity at the firm level using a production function approach, but also taking into account the impact of R&D spillovers. He finds a significant impact of R&D on firm performance on average. A meta-analysis on the studies surveyed shows that the estimated rates of return do not significantly differ among countries, whereas the estimated elasticities do. Furthermore, the estimated elasticities are significantly higher in the 1980s and consistently higher in the 1990s compared with the 1970s.

3.2 Interaction with the Demand Side

When equations (18.10) and (18.11) are estimated using firm data with sectoral deflators, the output measure (deflated nominal sales) also captures the impact of product innovation. Klette and Griliches (1996) deal with the problem of unobservable output prices including a demand equation. They consider that with an appropriate specification of the demand system, there is identification: the omitted price variable can be expressed in terms of the firm's output growth relative to industry output, and eventually in terms of observables and parameters already present in the production (or cost) function.¹⁵ In this regard, the supply-side approach can be improved by considering a demand function as follows:

$$Y_{it} = Y(P_{it}, G_{it}, D_{it}), \quad (18.14)$$

¹⁴ Klette (1996), following Hall (1988), develops a model to take into account the possibility of the existence of scale economies in addition to imperfect competition in the product market (mark-up different from one) considered by Hall (1988). He proposes an expanded equation of (18.10), which allows estimating the mark-up and the scale elasticity.

¹⁵ Nevertheless, Mairesse and Jaumandreu (2005) do not find big differences estimating the revenue function or the production function (using a real output measure) for French and Spanish firms. They point out that biases due to other sources of specification errors are probably more important.

where Y denotes the quantity of demanded output and P , G and D refer to price, knowledge capital and other demand shifters, respectively.

Technological capital affects firms' demand through the improvement of product quality or product innovations. The introduction of a new product on the market generates a new source of demand. The new product can replace the old product and in this case can cannibalize the revenues and the profits made from producing the old products. However, the new product can also be complementary to the old product. In this case, selling both products can generate economies of scale in the distribution of the goods on the market.

Assuming a Cobb-Douglas specification for equation (18.14) and taking log differences, we obtain:

$$\Delta y_{it} = \eta \Delta p_{it} + \xi \Delta g_{it} + \Delta d_{it}, \quad (18.15)$$

where lower-case letters denote logarithms of the variables and Δ denotes first differences of the variables. As in Klette and Griliches (1996), Klette (1996) uses the relationship between output and sales (revenue), $S_{it} = p_{it} Y_{it}$ to rewrite the TFP equation in terms of revenue rather than real output, under the assumption of an isoelastic demand equation. Klette (1996) assumes that each firm produces differentiated products and therefore faces its own downward-sloping demand curve, that the number of firms is large and that oligopolistic aspects of price-setting behavior are negligible. Because firms have idiosyncratic output prices, real revenue instead of an actual output measure is generated when revenue is deflated by an industry deflator. Taking log differences in the sales, $\Delta s_{it} = \Delta p_{it} + y_{it}$, and substituting in equation (18.15), the demand equation can be rewritten as:

$$\Delta y_{it} = \frac{\eta}{1+\eta} \Delta s_{it} + \frac{\xi}{1+\eta} \Delta g_{it} + \frac{1}{1+\eta} \Delta d_{it} \quad (18.16)$$

where $\frac{\eta}{1+\eta} = \mu$ is the price–marginal cost mark-up.¹⁶ Combining (18.11) and (18.16), the revenue production function can be expressed as:

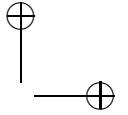
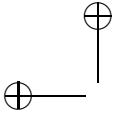
$$\Delta s_{it} = \frac{\Delta a_{it}}{\mu} + \frac{\alpha_l}{\mu} \Delta l_{it} + \frac{\alpha_k}{\mu} \Delta k_{it} + \frac{\alpha_m}{\mu} \Delta m_{it} + \left(\frac{\gamma}{\mu} - \frac{\xi}{\eta} \right) \Delta g_{it} - \frac{1}{\eta} \Delta d_{it} + \frac{\Delta u_{it}}{\mu}. \quad (18.17)$$

Equation (18.17) combines supply and demand effects, which implies that the parameter associated with R&D capital captures the effects of process innovations (cost reduction) and product innovations (demand increase).¹⁷ Because demand elasticity is negative, the parameter associated with this variable is positive. However, as the dependent variable is the revenue, the R&D elasticity is a combination of output and price elasticity. Identification is only possible with individual prices that allow us to estimate η and ξ separately.¹⁸

¹⁶ Alternatively, the demand equation can also be expressed in levels.

¹⁷ This equation can also be estimated in levels.

¹⁸ Van Leeuwen (2002) estimates this equation and, apart from including process innovation, he parameterizes the demand shifter by the share of new (or new and improved) sales in total sales. The basic assumption of this model is that innovation is predominantly “demand driven”, and therefore its contribution to productivity growth should be measured according to quality or product variety.



There are only a few papers that incorporate the demand side. In these studies, the impact of R&D on revenues is usually higher than the impact on physical output. See Hall et al. (2010) for a review of this literature.

3.3 The CDM Model

For the last two decades, the most commonly used model to analyze the effect of innovation on productivity has been proposed by Crépon et al. (1998), subsequently known as the CDM model. It takes into account the fact that not innovation inputs but innovation outputs increase productivity. Firms decide to invest in R&D in order to obtain some innovation outputs (product, process, organizational innovations, patents, etc.) that positively affect their productivity (or productivity growth) and other performance variables. In this regard, this model consists of a recursive system of three sets of equations. The first set of equations describes whether a firm undertakes R&D and, if so, how much, as a function of firm and industry characteristics.¹⁹ The second one takes the form of a knowledge production function, that is, explains innovation outcomes as a function of R&D intensity and other firm/industry characteristics. Finally, a productivity equation is considered where innovation outputs are factors among other inputs.

The first set explains the probability of undertaking R&D and the intensity of the R&D expenditure. The R&D effort R_{it}^* can be measured by the intensity of the R&D expenditure R_{it} only if the firm makes (and reports) that expenditure. The decision to perform R&D expenditures is represented by:

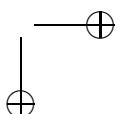
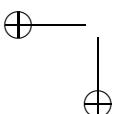
$$dR_{it} = \begin{cases} 1 & \text{if } dR_{it}^* = x_{1it}'\beta_1 + \varepsilon_{1it} > 0, \\ 0 & \text{if } dR_{it}^* = x_{1it}'\beta_1 + \varepsilon_{1it} \leq 0 \end{cases}, \quad (18.18)$$

where dR_{it} is a binary variable that takes the value 1 when the firm invests in R&D, and 0 otherwise. If the latent variable dR_{it}^* is bigger than a constant threshold (which can be zero), we then observe that the firm engages in R&D activities. x_{1it} is a vector of observable explanatory variables (time-variant and time-invariant variables). Conditional on the performance (and report) of R&D activities, the second equation refers to the quantity of resources allocated to this purpose:

$$R_{it} = \begin{cases} R_{it}^* = x_{2it}'\beta_2 + \varepsilon_{2it} & \text{if } dR_{it} = 1 \\ 0 & \text{if } dR_{it} = 0 \end{cases}, \quad (18.19)$$

where x_{2it} is a vector of determinants of the innovative effort, which can differ from those determinants that explain the decision to perform R&D expenditures. Finally, ε_{1it} and ε_{2it} are idiosyncratic errors (which refer to other unobservable time-variant determinants). Most papers assume that the error terms ε_{1it} and ε_{2it} follow a bivariate normal distribution with a mean equal to 0, variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1$, and correlation coefficient ρ_{12} . For this reason, both equations are usually estimated as a generalized Tobit model by maximum likelihood

¹⁹ Sometimes this first block involves just one equation when the sample refers to innovative firms.



(Heckman, 1976, 1979). In the original paper, the innovative effort is approached by the research-technological capital per employee. Posterior empirical evidence usually considers the R&D expenditure per employee a proxy of technological effort.

The third equation of the model corresponds to the estimation of the new knowledge production function, g_{it} , generated from firms' innovative effort. The model assumes that the investment intensity is a public good within the firm that can be used to produce different outputs without depletion. Therefore, g_{it} can be modeled as a vector of technological outputs that can take several forms:

$$g_{it} = \lambda R_{it}^* + x_{3it}'\beta_3 + \varepsilon_{3it}, \quad (18.20)$$

where latent investment intensity R_{it}^* appears as an explanatory variable with the vector x_{3it} , which includes other determinants of knowledge production (time-variant and time-invariant variables).

In the original paper, new knowledge is measured by two variables: the number of patents and the percentage share of firm innovative sales (products launched in the market in the last five years). In the first case, the patent equation is specified as a heterogeneous count data process. In the second one, since the share is only known by intervals, the equation is specified as an ordered probit model. The choice of the dependent variable is conditioned by the availability of data in Community Innovation Surveys (CIS), which are the databases most frequently used to estimate the CDM model with European data. Subsequent studies complement these indicators with dummy variables that capture the achievement of product and process innovation and, recently, organizational innovation.

Finally, the last equation refers to the productivity equation. Most papers consider an augmented Cobb-Douglas production function with physical capital, employment, innovation outputs and other control variables. In addition, they usually assume constant returns to scale (in standard inputs). The functional form that has been used more frequently following the original CDM is:

$$y_{it} - l_{it} = \delta g_{it} + x_{4it}'\beta_4 + \varepsilon_{4it}, \quad (18.21)$$

where $y_{it} - l_{it}$ is labor productivity (added value or output per worker, expressed in log), innovative output g_{it} appears as an explanatory variable with the vector x_{4it} , which includes other determinants like physical capital per employee (in log) and skill composition, among others. However, recent empirical evidence considers other augmented Cobb-Douglas production functions like equations (18.10) and (18.17) of the previous section.²⁰

In the original paper (Crépon et al., 1998) all equations ((18.18)–(18.21)) are estimated jointly by asymptotic least squares or a minimum distance estimator. In the first stage, they estimate the reduced-form equations parameters by M-estimation, and in the second stage, ALS-estimation to retrieve consistent estimates of structural parameters. However, successive empirical evidence uses a sequential approach: the predicted value of the dependent variable of each set enters as a determinant in the next equation.

²⁰ Klomp and Van Leeuwen (2001) also have exploited the CIS data in a structural modeling approach. But in contrast to Crépon et al. (1998), they do not use a production function framework.

Specifically, after the estimation of equations (18.18) and (18.19) of the first set, the predicted value for all firms is used as a proxy of the innovation effort and included as a determinant in equation (18.12), the knowledge production function. This implies that the model assumes that all firms make some innovative effort even if they do not report this effort. That is, below a certain threshold, the firm is not capable of picking up explicit information about this effort and will not report on it. In this regard, equation (18.20) is estimated for all firms and not only for the subsample of those reporting R&D expenditures. In addition, the predicted value instead of the observed value dR_{it}^* is used as an explanatory variable to take into account the potential endogeneity of this variable in the knowledge production function: some unobservable characteristics of firms can generate both innovative effort and technological output increases. In the last step, the productivity equation is estimated by using the predicted values of g_{it} to take care of the endogeneity of this variable in equation (18.21). In this regard, the CDM model takes into account the endogeneity of R&D effort and innovation outputs in the knowledge and productivity equation, respectively. However, there is no feedback from productivity to innovative activities.

Revisions of this literature have been provided by Hall (2011) and Mohnen and Hall (2013). The empirical evidence shows that productivity is positively related to both innovative sales and the binary indicator for product innovations and that the association is higher for high-technology sectors.²¹ However, regarding process innovation, results vary a lot and, in some cases, are even negative. Two possible explanations have been provided (Hall, 2011): (i) firms operate in the inelastic portion of their demand curve so that revenue productivity is not affected by efficiency improvements in the production process; and (ii) there is so much measurement error in the innovation variables that only one of the two is positive and significant when added to the productivity equation.²²

3.4 Persistence in Technological Inputs and Outputs

Although there is ample empirical evidence that tests the CDM model, most of it is based on cross-sectional data. This fact does not permit the estimations to take into account persistence in R&D activities, dynamic linkages between innovation and economic performance or unobserved firm heterogeneity. Recently, some studies have begun to consider the timing of innovation and dynamics aspects.

There are papers that show persistence in economic performance (profits and productivity) of firms. For example, Cefis and Ciccarelli (2005) find long-run persistence in profit differentials among UK manufacturing firms. In addition, their results show a positive difference in profitability between innovators (firms that apply for patents) and non-innovators, which is greater when the comparison is between persistent innovators and non-innovators. Bartelsman and Dhrymes (1998) and Fariñas and Ruano (2005) give evidence of persistence in the differences in productivity across firms for US and Spanish manufacturing firms, respectively.

²¹ The effect tends to be lower when growth rather than level of productivity is estimated and when skilled labor is controlled for, suggesting an identification problem between innovation and other measures of knowledge and physical capital.

²² Van Leeuwen and Klomp (2006) present a structural model similar to the CDM model but they use revenue per employee growth as the measure of firm performance instead of the level of value added per employee in the performance equation (equation (18.17) of Section 3.2).

The static version of the CDM model shows that output innovation is related to productivity and, in this regard, persistence in innovation activity can explain the persistence in firm economic performance.

More recently, the availability of new data allows estimating the growth of labor productivity or TFP by using panel data and introducing dynamics in the CDM model or, at least, in the equations for the decisions on technological inputs or outputs. There are two main explanations for persistent behavior: true state dependence and spurious dependence. The first one implies a real causal effect: the probability of investing in $t - 1$ increases the probability of investing in t . Explanations for this real true dependence in the case of innovation activities are the sunk costs associated with the performance of R&D activities, the "success breeds success" hypothesis and the existence of dynamic increasing returns, among others (see, for example, Peters, 2009; Mañez-Castillejo et al., 2009; and Raymond et al., 2010).

Sunk costs represent a barrier to both entry into and exiting from R&D activities. If a firm decides to undertake R&D investment, it has to incur start-up expenditures to build an R&D department. These costs are an entry barrier because potential entrants have to take them into account in their profit-maximization behavior. They are also a barrier to exiting because they are not recovered when the firm stops R&D activity and it has to incur them (maybe in a smaller quantity) again if it decides to re-enter.

Another explanation is the "success breeds success" hypothesis, which is based on different arguments in the literature. First, and following Schumpeter, if there is a positive relationship between market power and innovation, incumbents have more to lose by not innovating than potential new entrants do and this causes incumbents to innovate persistently. In addition, a firm's innovation success broadens its technological opportunities, which make subsequent innovation success more likely. Another argument is the existence of financial constraints. Innovation projects entail a high risk and, because of information asymmetry between the innovator and the lender, firms have problems obtaining external funds. Profits that are generated by past successful innovations provide firms with increased internal funding that can be used to finance further innovations.

The last explanation, as Peters (2009) points out, is based on the idea that knowledge accumulates over time (Nelson and Winter, 1982). Evolutionary theory states that technological capabilities are a decisive factor in explaining innovation. Experience in innovation is associated with dynamic increasing returns in the form of learning effects that enhance knowledge stocks and, hence, technological capabilities. In addition, evolutionary theory defines the notion of technological trajectory. Along a trajectory, radical innovations are followed by a succession of incremental innovations. Consumers are inclined to buy new generations of products, increasing the demand for innovation.

According to these theoretical explanations for real state dependence, it is not clear whether persistence is more related to technological inputs or outputs. Under the sunk cost hypothesis, R&D decisions are modeled on a long-term horizon, given that sunk costs could represent a barrier not only to entry for new firms, but also a barrier to exiting for incumbent firms that have not recovered their investments. In this case, an input measure would be desirable. However, the "success breeds success" and the "learning by doing" hypotheses are more associated with technological results. Additionally, if we assume that innovation outputs are in part determined by innovation inputs, input persistence should be partially translated into output persistence.

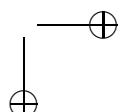
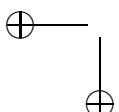
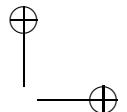
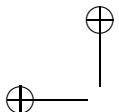
From an empirical point of view, there is a lower number of papers focused on innovation inputs (R&D expenditure) than on technological outputs (patents or process and product innovations). As for the first, two outstanding studies are the ones developed by Mañez-Castillejo et al. (2009) and Peters (2009). Both obtain evidence in favor of a high degree of persistence regardless of the methodology. In particular, Mañez-Castillejo et al. (2009) estimate a multivariate dynamic discrete choice model of R&D decisions using firm-level data of Spanish manufacturing for 1990–2000. Conditional on firm heterogeneity and serially correlated unobservable factors, they find that R&D history matters. They interpret this true dependence as the existence of sunk R&D costs associated with performing R&D. They deal with econometric problems (error term serially correlated because of permanent firm-specific component and the initial conditions problem) following Heckman (1981).²³

Peters (2009) analyzes the persistence of firms' innovation for German manufacturing and service firms for the period 1994–2002. She estimates a dynamic random effects discrete choice model and uses the estimator proposed by Wooldridge (2005). The econometric results show that past innovation experience is an important determinant for manufacturing and for service sector firms alike, and hence confirm the hypothesis of true state dependence. She defines an innovator as a firm with positive innovation expenditure in a given year. In this regard, the paper analyzes persistence in innovation input.

With respect to innovation outputs, apart from patents, the empirical evidence is also consistent with the existence of high persistence. Duguet and Monjon (2004) show that persistence in (process or product) innovation is strong for a sample of French manufacturing firms. Using data on Australian firms, Rogers (2004) also finds that there is some degree of persistence in innovative activities defined in terms of product and organizational innovations. However, neither study controls for unobserved individual heterogeneity. Flaig and Stadler (1994) deal with this problem and examine persistence in process and product innovations using a panel of manufacturing firms in West Germany in the 1980s. They estimate a dynamic panel probit model that accounts for unobserved firm-specific characteristics following Heckman's (1981) approach. For both types of innovation, their results suggest the existence of true state dependence, which implies a positive impact of innovative success on further innovations in the following years.

In the case of patents, most papers find a low level of persistence (see, Malerba and Orsenigo, 1999; Cefis and Orsenigo, 2001; and Cefis, 2003). This result is confirmed by Geroski, Van Reenen and Walters (1997) using a duration model for granted patents for a sample of US firms to test the "success breeds success" idea: very few innovative firms are persistently innovative. This low persistence when innovation activity is measured with patents can be explained because not all inventions are patented. In addition, as Duguet and Monjon (2004) point out, patenting involves both innovating and being the first to innovate. In this regard, patent data could measure the persistence of innovative leadership rather than the persistence of innovation. With a panel of French manufacturing firms, Crépon and Duguet (1997) also use patent data to measure innovation. They estimate a dynamic count data model that links the current number of patents to both the previous year's number of patents and the

²³ Heckman (1981) suggests starting on the joint distribution of the dependent variable in all periods conditioned to the permanent component of the residual and the mean of the explanatory variables. He also proposes specifying the distributions of the initial condition and the permanent component to the error term to integrate out the unobserved effect.



amount invested in R&D. They find that the effect of lagged patents on the current number of patents is significantly positive, which suggests a persistence in innovation among formal R&D performers, but the effects slowly vanish as the number of innovations increases.

Previous empirical evidence analyzes persistence in the R&D activities in just a single equation without considering the links among technological inputs and outputs and productivity. Recent papers provide evidence about the relevance of taking persistence into account when examining these links. In particular, Raymond et al. (2009) consider the possibility of dynamics in the first two equations of the CDM model for Dutch manufacturing, using an unbalanced panel from five waves of the CIS during 1994–2004. They consider not only the persistence of innovation input and innovation output, but also the lag effect of innovation input on innovation output and the feedback effect of innovation output on innovation input.²⁴ They find persistence of innovation input and innovation output, a lag effect of the former on the latter and a feedback effect of the latter on the former.

Van Leeuwen (2002) analyzes the dynamics of R&D intensity and links it to that of innovative sales using two waves of the Dutch CIS data. The results show that innovation persistence is smaller when measured from the output side than when judged from R&D intensities. As in Van Leeuwen and Klomp (2006), he also includes a productivity equation in terms of revenue growth. In this last equation, the returns of the current R&D endeavor are more significant if the dynamic specification is relaxed, and a restricted and static model is applied to all available CIS data.

Huergo and Moreno (2011) specifically take into account persistence in R&D activities and technological outputs inside the CDM model for a sample of Spanish manufacturing firms between 1990 and 2005. For the first two equations, they estimate a Heckman model with a dynamic pooled probit for the decision whether to engage in R&D activities or not, where the individual heterogeneity is parameterized as in Wooldridge (2005). In the third equation (new knowledge generation), the lagged dependent variable is added as an explanatory variable to reflect whether the firm has previously generated new knowledge capturing the innovation output experience. Because of the binary character of innovation measures, this equation is estimated as a random effect (RE) dynamic probit model. The last equation of this paper is a TFP growth equation, where TFP is calculated as a Solow residual, imposing the usual assumptions. In the estimation, the authors assume a recursive model where feedback from productivity growth to technological effort is not allowed, and therefore a three-stage estimation procedure is applied. The results reflect the existence of true state dependence both in the decision of R&D investment and in the production of innovations. The omission of this persistence leads to an overestimation of the current impact of innovations on productivity growth.

The previous paper does not take into account the potential correlation of the error terms across equations or the dynamics in the productivity equation. Raymond et al. (2015) deal with these challenges and introduce dynamics in the R&D-to-innovation and innovation-to-productivity relationships. They consider four non-linear dynamic simultaneous equations and estimate them by full information maximum likelihood using two unbalanced panels of Dutch

²⁴ To do so, a dynamic panel data bivariate Tobit with double-index sample selection accounting for individual effects is estimated by maximum likelihood.

and French manufacturing firms from three waves of the CIS.²⁵ The results provide evidence of robust unidirectional causality from innovation to productivity and of stronger persistence in productivity than in innovation. Specifically, they only find true persistence in product innovation in French manufacturing.²⁶

Using an alternative approach, Deschryvere (2014) analyzes what role persistence of innovation output plays in the growth of firms. Specifically, he applies a vector autoregression model to Finnish firm-level data. He finds that only continuous product and process innovators show positive associations between R&D growth and sales growth. In the case of process innovations, occasional innovators are the ones that exhibit a stronger association between sales growth and subsequent R&D growth.

3.5 Recent Structural Models

Previous empirical analyses on the effect of R&D at the firm level are based on the estimation of the relationship between technological input (or technological outputs as in the CDM model) and the level of output. The marginal product of the knowledge input (technological capital) in the production function provides a measure of the return of the firm's R&D expenditure.

An alternative way to incorporate R&D in the firm's production process has been developed by Aw et al. (2011) and Doraszelski and Jaumandreu (2013). Instead of building a knowledge capital, like Griliches (1979), they assume that productivity is unobservable and model it as a Markov process that depends on the R&D expenditure. That is, productivity is affected by the firm's endogenous decision to invest in R&D. In the case of Aw et al. (2011), the export choice also endogenously affects the path of productivity.²⁷ Specifically, from the standard Cobb-Douglas production function in logarithms:

$$y_{it} = a_0 + \alpha_t t + \alpha_l l_{it} + \alpha_k k_{it} + \alpha_m m_{it} + w_{it} + u_{it} \quad (18.22)$$

Doraszelski and Jaumandreu (2013) assume that actual productivity, w_{it} , can be decomposed into expected productivity and a random shock:²⁸

$$w_{it} = E[w_{it}|w_{it-1}, R_{it-1}] + \xi_{it} = G(w_{it-1}, R_{it-1}) + \xi_{it}, \quad (18.23)$$

where R_{it-1} is R&D expenditures in $t - 1$. That is, the firm anticipates the effect on R&D productivity in period t when making the decision about investment in knowledge in period $t - 1$. Using an unbalanced panel of Spanish manufacturing firms during the 1990s, they found that R&D expenditures determine the differences in productivity across firms and

²⁵ Specifically, they take care of the initial conditions problem using Wooldridge's (2005) approach and they handle multiple integration due to the correlations of firm effects and idiosyncratic errors across equations by using Gauss-Hermite quadrature sequentially (Raymond et al., 2007).

²⁶ In a previous paper, using the same estimation technique, Raymond et al. (2010) study the persistence of product or process innovation and of innovative sales. They estimate a dynamic Type II Tobit model and find that the intensity of past innovation output affects current intensity in high-tech activities, but there is no effect in low-tech activities.

²⁷ To deal with the endogeneity, following Olley and Pakes (1996), the econometric literature models productivity as an exogenous first-order Markov process.

²⁸ As in Olley and Pakes (1996), the conditional expectation function $G(\cdot)$ is not observed by the econometrician and must be estimated non-parametrically along with the parameters of the production function.

the evolution of firm-level productivity over time. They also provide evidence of the non-linearities and uncertainty in the R&D process.

Aw et al. (2011) use a model of firm revenue in domestic and export market instead, but instead of using equation (18.22) from a standard production function, they consider a short-run marginal cost function and firms operating in monopolistic competitive markets (domestic and foreign) that apply a mark-up to the marginal cost. The firm's revenue in each market depends on the aggregate market conditions, the capital stock, the vector of variable inputs prices and the firm-specific productivity that they model as:

$$w_{it} = G(w_{it-1}, dR_{it-1}, dX_{it-1}) + \xi_{it}, \quad (18.24)$$

where dR_{it-1} and dX_{it-1} , are the firm's R&D and export market participation in $t - 1$, respectively. As in Doraszelski and Jaumandreu (2013), they assume that the firm can affect the evolution of its productivity by investing in R&D, but now they also consider the possibility of learning-by-exporting. That is, being an exporter is a source of knowledge and can improve future productivity. Using plant-level data for the Taiwanese electronics industry, they obtain that both activities, R&D and export, positively affect the future productivity of the plants. Nevertheless, R&D marginal effect varies a lot with productivity; it is much higher for high-productivity plants, and it is higher for non-exporting firms (although the magnitudes of these second results are quite small).

Some recent articles complement the previous papers and take a different approach to deal with the simultaneity problem by modeling and estimating the firm's dynamic decision to invest in R&D instead of estimating just a production function (Roberts and Vuong, 2013; Peters et al., 2013). They elaborate a dynamic structural model of R&D demand so that the expected benefit to the firm's investment in R&D is inferred from the rational decision of R&D investment. One important advantage of these models is that they allow the change of parameters in the firm environment²⁹ and the quantification of the effect of these changes on the firm's decision to invest in R&D, productivity and the long-run impact on profitability.

The antecedents of these models can be found in Rust's (1987) model for discrete investment decision by firms and have been applied to a wide range of a firm's choices, for example, export decisions in international trade (Das, Mark and Tybout, 2007) or entry and exit choices in industrial organization (Aguirregabiria and Mira, 2007; Collard-Wexler, 2013).

The starting point of the dynamic structural models for R&D demand is that R&D investment is an inherently dynamic decision because the firm must incur costs in the present period for an anticipated gain in profits in future periods. This gain shows several important features: (i) there is a time lag between investment and output; (ii) R&D is unlikely to have a one-time impact; and (iii) the magnitude of the gains is surrounded by uncertainty. Structural models accounts for these features.

We will present a summarized version of this kind of model, following Robert and Vuong (2013).³⁰ A more complex version can be found in Peters et al. (2013). They consider a single firm, i that makes input choices at the beginning of time period t and faces a logarithmic production function as equation (18.22). As in previous papers, they consider w_{it} to be a firm-specific productivity level that the firm observes and u_{it} a random shock that the firm does

²⁹ For example, the degree of competition in the output market or the introduction of R&D subsidies.

³⁰ In their paper, Robert and Vuong (2013) relate the equations of their model to the CDM model.

not control and does not observe in advance. In the simplest version, the firm chooses l (and other variable inputs) that maximizes the profit function, $\pi(w_{it})$. The important point here is that the future firm productivity level (w_{it}) directly impacts profit and can be affected through R&D investments.³¹

Accordingly, the firm must decide whether to invest in R&D (R_{it}) to improve the level of its future productivity. This decision is based on the comparison of costs and benefits of R&D. On the one hand, costs of R&D, $c(r_{it})$ include not only expenditures on R&D inputs but also adjustment costs. Adjustment costs will be higher for those firms not performing R&D in the previous period than for those firms already developing R&D activities, reflecting the existence of an entry cost to R&D. It is also acknowledged that some R&D costs (such as capital costs) will not be observed by the research and should be estimated within the model.

On the other hand, the benefits of R&D are treated in two steps. First, the firm's choice of R&D affects the probability of realizing an innovation in the next period, $F(g_{it+1}|R_{it})$, where n_{t+1} stands for innovations in $t + 1$. This function recognizes that there is uncertainty in the innovation process, so some R&D efforts may fail. In addition, it allows innovation without any formal R&D spending.³² In the Peter et al. (2013) paper, R_{it} is measured as a discrete variable, dR_{it} , and the innovations, g_{it+1} , are also defined as discrete variables.

Second, innovations can lead to improvements in the firm's future productivity, which is represented by a distribution function $G(w_{it+1}|w_{it}, n_{it+1})$, that depends on both the firm's current productivity and its realized innovation.³³ This specification recognizes the persistence of firm-level productivity and that innovations have a lasting effect on productivity:

$$w_{it} = G(w_{it}, g_{it+1}) + \xi_{it+1}. \quad (18.25)$$

That is, Peters et al. (2013) and Roberts and Vuong (2013) include the innovation process in the model (the first step) instead of linking R&D to productivity like Aw et al. (2011) and Doraszelski and Jaumandreu (2013). Combining both steps, they capture not only the endogeneity of the productivity process but also the uncertainty of the innovation process. In addition, that allows them to analyze whether R&D expenditure improves productivity through the demand side (product innovations) or cost side of the firm's operations (process innovations). These features of the model allow for an explicit formulation of the firm's dynamic demand for R&D that is absent from the knowledge capital or CDM model.

The firm i chooses its sequence of R&D expenditures (R_{it}) to maximize the discounted sum of expected future profits net of the costs of R&D. Its value function can be written as:

$$V(w_{it}) = \pi(w_{it}) + \max_{r_{it}} [\beta E V(w_{it+1}|w_{it}, R_{it}) - c(R_{it})], \quad (18.26)$$

³¹ The profit function will also depend on the fixed inputs (k) and the exogenous input prices (the firms apply a mark-up on the short marginal cost like Aw et al., 2011). The more complex model incorporates exogenous state variables in the specification (Peters et al., 2013) and explicit equations for firm cost and demand, so that the profit function is derived from them.

³² Innovation without R&D is quite widespread and it could be driven by other innovation inputs, such as design, training or hiring of new people.

³³ This is analogous to the third equation in the CDM model.

where β is the discount factor, and the firm's value function, V , is the sum of the current period profit, π , and the maximized discounted future expected value net of the cost of investment. The term $EV(w_{it+1}|w_{it}, R_{it})$ is important because it captures all future payoffs to the firm from R&D investment and it can be written as:

$$EV(w_{it+1}|w_{it}, R_{it}) = \int_{n,w} V(w_{it+1}) dG(w_{it+1}|w_{it}, g_{it+1}) dF(g_{it+1}|R_{it}), \quad (18.27)$$

where the three terms on the right-hand side of the equation identify the three steps: (i) from R&D to innovation; (ii) from innovation to future productivity; and (iii) from future productivity to future long-run profits.

As previously mentioned, empirical papers have focused on the discrete choice of R&D, so the R&D variable is given as $dR_{it} = 1$ if the firm invests in R&D and $dR_{it} = 0$ if it does not. Therefore, the long-run payoff of R&D is defined as:

$$\Delta EV(w_{it}) = EV(w_{it+1}|w_{it}, dR_{it} = 1) - EV(w_{it+1}|w_{it}, dR_{it} = 0), \quad (18.28)$$

The term ΔEV is simply the increment to the expected (long-term) future value of the firm if it chooses to invest in R&D in period t .³⁴

When deciding to invest in R&D the firm weighs this expected gain in future profits against the current cost of R&D so that the final element to complete the model is the specification of the cost function for R&D. Like Aw et al. (2011), Peters et al. (2013) treat firm costs as independent draws, ψ_{it} , from an underlying cost distribution, $C(\psi)$, and estimate parameters describing this distribution as part of the model.

In the discrete framework, the firm's demand for R&D is simply the probability that they choose to invest in R&D, that is, the probability that $\Delta EV \geq \psi_{it}$.

Peters et al. (2013) estimate the model using firm-level data from the Manheim Innovation Panel for German manufacturing results. The main results show that, expressed as a proportion of firm value, the net benefit for the median firm with prior R&D experience varies from 2.4 to 3.2 percent across five high-tech industries but varies from -4.6 to 0.6 percent for firms without previous R&D experience. This negative value implies that the median inexperienced firm would not find it profitable to invest in R&D. Given inexperienced firms find R&D profitable, the net benefit of starting R&D varies from 2.0 to 2.4 percent of firm value. In low-tech industries the net benefits are substantially smaller. In addition, they simulate how changes in the R&D cost affect the firm's choice and, consequently, future productivity. For example, in high-tech industries a 20 percent reduction in fixed R&D cost leads after five years to an average increase of 7 percentage points in the probability of investing in R&D and a 4 percent increase in productivity.

Robert and Vuong (2013) estimate a simplified version of the model estimated in Peters et al. (2013) for German data. They find similar results: the expected benefit of R&D investment varies positively with the firm's productivity, and is substantially larger in a group of high-tech industries than in a group of less R&D-intensive industries. As a share of a firm's value, the expected benefit of R&D net of R&D costs varies from -0.6 percent to 1.6 percent

³⁴ A higher level of productivity would imply a higher return on the investment, thus making it more likely that the firm will invest in R&D in the future.

across firms with different productivity levels in high-tech industries, and from -3.3 percent to 0.8 percent in low-tech industries. Firms with negative net benefits would choose not to invest in R&D.

4 INSIDE THE R&D BLACK BOX

The literature reviewed so far is generally constrained by a lack of information on and analysis of R&D specificities. This section is aimed at summarizing two strands of literature that seek to open up and examine the contents of the R&D black box. First, there is literature that has attempted to open up the R&D black box by explicitly taking into account that research and development are two different activities, and therefore may differ in terms of their determinants and effects. Second, firms often adopt a number of innovation strategies simultaneously, and this coexistence of strategies suggests the existence of complementarities. Here we focus on empirical studies aimed at identifying complementarity relationships in R&D activities.

4.1 R&D Composition

One limitation of previous literature is that R&D has been assumed to be a homogeneous activity. However, research and development are two different activities that differ in purposes, knowledge bases, the people involved and management styles (Barge-Gil and López, 2015). More precisely, the main purpose of research is to acquire new knowledge, while the main purpose of development is directed towards the introduction of new or improved products or processes (OECD/Eurostat, 2005). Research is more theoretical in nature (although frequently oriented to some practical objective) and is based on analytical knowledge. Development is essentially applied and based on synthetic knowledge (Asheim and Coenen, 2005). Research needs specialized human capital that works relatively independently of the rest of the organization and without much hierarchy, while development shows clear hierarchy and needs generalists able to coordinate with other functions of the organization (Leifer and Triscari, 1987). As a consequence, research and development are increasingly carried out in different departments, even located in distant places (Chiesa, 2001).

The issue of heterogeneity in R&D was addressed in seminal works by Mansfield (1980, 1981), Link (1981, 1982, 1985), Griliches (1986) and Lichtenberg and Siegel (1991). However, these authors themselves point out the limitations of their studies and stress that their results should be viewed as preliminary. The reason is that they used small samples of very large US firms and usually were not able to address issues of simultaneity and endogeneity. Two main topics were addressed: R&D determinants and R&D effect.

Regarding the first topic, Mansfield (1981) uses a survey of 108 large US firms to analyze the determinants of the composition of R&D expenditures and the effect of this composition on innovative output. Four types of R&D expenditures are distinguished: (i) devoted to basic research; (ii) devoted to relatively long-term projects (five or more years); (iii) aimed at entirely new products and processes; and (iv) devoted to relatively risky projects (less than a 50–50 estimated chance of success). He finds that these four dimensions of R&D are not related much (when comparing firms within industries) and that larger firms are more oriented towards basic research. In addition, there is some correlation between the number

of innovations and the proportion of basic research on total R&D expenditures. Link (1982) analyzes the determinants of basic research, applied research and development for a sample of 275 firms belonging to the Fortune 1000 list in the USA. He finds that orientation to development is higher for firms operating in more concentrated markets and receiving more public funding, while firms with a higher level of profits are more oriented to applied research. Finally, orientation to basic research increases with diversification and profits and was higher for owner-managed firms. In a later work, Link (1985) adopts a dynamic perspective. He finds that orientation to basic and long-term research is decreasing, so he analyzes the determinants of this change for 146 very large US firms. He finds that managerial issues are important as firms with a more offensive strategy and central R&D labs are also those more increasingly oriented towards basic and long-term research.

Regarding the second topic, Mansfield (1980) uses data from 119 US firms to analyze the effect of R&D composition on productivity. He finds that there is a positive relationship between the amount of basic research and productivity, after holding constant other R&D expenditures (which do not show an effect on productivity). Griliches (1986) analyzes the relationship between R&D and productivity growth in approximately 1000 large US firms from 1957 to 1977. He finds that basic research appears to be more important as a productivity determinant than other types of R&D and that privately financed R&D expenditures are more effective than federally financed ones.³⁵ These results hold when individual firm effects are removed. Finally, Lichtenberg and Siegel (1991) use an improved dataset, with more than 2000 firms (including some small ones) accounting for 84 percent of the R&D performed in the USA in 1976. The data allow them to control for firm diversification so that efficiency of estimation is improved. They find that only investment in basic research shows a positive effect on productivity (neither applied research nor development show effects different from zero) and that company-funded R&D shows a positive effect on productivity, but not federally funded R&D.³⁶

However, to our knowledge, in spite of the relevance of these papers and claims by their authors about the importance of studying the composition of R&D, this topic has not received much attention for a long period. In the last few years, however, interest has been renewed, driven by the availability of new data from CIS surveys. Three main topics have been addressed: the relationship between public funding and the composition of R&D, the different determinants of R&D components and the different effect of these components on innovation outputs.

Regarding the first topic, Aerts and Thorwarth (2009) use a sample of 521 Flemish firms from two waves of the R&D survey (2004 and 2006) and find that additionality of public funding exists in development but not in research. Clausen (2009) uses a sample of 1019 firms in Norway and distinguishes between subsidies for research and subsidies for development and finds that there is additionality for research but not for development subsidies. Czarnitzki, Hottenrott and Thorwarth (2011) use an unbalanced panel (1999–2007) including 952 Belgian firms. These authors analyze financial constraints associated with R&D activities. They find a higher effect of financial constraints on research (this activity is

³⁵ Confirming at the firm level, the results obtained at the industry level by Griliches and Lichtenberg (1984).

³⁶ In an appendix to their work, Hall and Mairesse (1995) explore the compositional effects of R&D in a sample of 197 firms during eight periods. They find that the fraction of R&D devoted to basic research reduces overall productivity by 5–9 percent (standard error of 2–3 percent) and government-funded research does not seem to have much effect until it raises over 20 percent of the firm's R&D budget. At this point, its effect is positive.

performed by firms showing more liquidity and less debt than those performing development). In a related work, Czarnitzki and Hottenrot (2011) use an unbalanced panel (1993–2002) of 352 German firms for ten years and distinguish between “routine” vs “cutting edge” R&D investment and conclude that financial constraints exist for “cutting edge” R&D but not for “routine” R&D.

Regarding the second topic,³⁷ Barge-Gil and López (2014) use a sample of more than 4000 firms per year in the period 2005–09, accounting for the correlation between error terms from a research and development equation. They find that demand pull and appropriability have a higher effect on development activities while technological opportunity has a higher effect on research activities. Additionally, for larger firms the effect of size is usually higher on development than on research and no important effect is achieved for market power. The work by Czarnitzki and Hottenrot (2011) controls for the Herfindahl index of market concentration and finds a positive effect of this index on routine R&D but not on cutting-edge R&D.

Regarding the third topic, Lim (2004) analyzes the different impact of basic and applied research in firms from pharmaceutical and semiconductor industries. His sample is composed of an unbalanced panel (1981–97) containing 1129 observations for the semiconductor industry and 571 for the pharmaceutical industry. He finds that applied research shows a much higher effect on the number of patents than basic research (actually, basic research shows a negative effect in the semiconductor industry). Czarnitzki, Kraft and Throwarth (2009) analyze the different impact of research and development on patents. They use an unbalanced panel of 122 Flemish firms from 1993 to 2003. They find that the patent-R&D relationship exhibits a premium for the portion of R in R&D although they warn about the explorative nature of the result due to the small size of the sample used. In a later work, Czarnitzki and Thorwall (2012) argue that previous studies do not control for technological opportunity and appropriability, so they analyze the different impact of basic research in low-tech and high-tech industry, using a sample of 353 Flemish firms observed in three periods. They find that there is a premium for investment in basic research in high-tech industries but no premium (or discount) exists in low-tech industries. Additionally, they find that the premium for basic research increases with firm size. They interpret these results as showing that appropriability conditions for basic research are lower in low-tech industries and in small firms. Finally, Barge-Gil and López (2015) use a sample of 4024 Spanish firms for the years 2005–08 to estimate the differentiated effect of research and development in several innovation outputs: patents, new products and processes and sales from new products. The main findings show that both activities contribute to each innovation output, but the contribution of research is higher for process innovation, while the contribution of development is higher for new products, and especially for sales from new products. On an industry level, they also find that research shows a greater effect of sales from new products in low-tech sectors.

³⁷ Some theoretical works have also been developed in this line. Cabral (1994) proposes that market power is associated with development (rather than research) activities in a static framework. In a later work, Cabral (2003) extends this result to a dynamic framework, using a model with a leader and a laggard. In this case, he finds that the optimal choices are pursuing safer (development) projects for leaders and pursuing riskier (research) projects for laggards. The model proposed by Kwon (2010) suggests that lower market power leads to riskier and longer-term projects (research).

4.2 Complementarity in R&D Activities

In its more general definition, complementarity between practices is understood to exist if the returns to adopting one practice are greater when the other practices are present. Literature on empirical industrial organization has long been interested in the analysis of complementarity between practices or decisions. Researchers in the field of economics of innovation, and in particular those interested in R&D activities, are not indifferent to this tendency.

The objectives of this section are twofold. First, we summarize the main econometric approaches used for testing for complementarity. Second, we review the main contributions to the empirical literature on complementarity in R&D activities.

4.2.1 Empirical methods for testing complementarity

The formal study of complementarity relies on analyzing the interactions among pairs of decisions and it can be traced back to Topkis (1978). Topkis (1978) formulated the concept of complementarity within the mathematical theory of lattices, while Milgrom and Roberts (1990) and Vives (1990) first applied this approach to economics. For a recent review of the precise mathematical treatment of complementarity, interested readers are referred to Brynjolfsson and Milgrom (2013). These authors summarize the theory about decision problems with complementarities in ten theorems based on the results by Topkis (1978), Milgrom and Roberts (1995, 1996), Milgrom and Shannon (1994), and Milgrom, Qian and Roberts (1991).

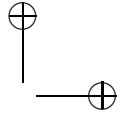
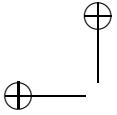
This section is focused on the empirical analysis of complementarity. Literature has proposed two principal ways in which complementarities reveal themselves empirically. First, complementary practices are often more likely to be adopted jointly rather than separately (i.e., clustering of practices across firms). Second, complementary practices are often more productive when adopted together than when adopted separately. These two empirical predictions are the basis for the statistical methods used to assess the existence of complementarities (the review of these different approaches done by Athey and Stern, 1998 remains a cornerstone of this literature).

There are two main approaches for testing the existence of complementarities. The first approach, the so-called adoption or correlation approach, consists of estimating correlations and demand equations. The second approach, the so-called productivity, production or performance approach, consists of estimating performance differences.

For simplicity in the exposition, in what follows, we consider a case with two potential complements, y_1 and y_2 . We will refer to the generalization to the case of more than two complements when necessary. Here, we closely follow the arguments and the notation used by Brynjolfsson and Milgrom (2013).

Adoption approach This approach is based on the *revealed preference principle*: under the assumption of optimizing behavior of the firm, the joint adoption of practices is potentially informative about the joint returns generated by them. This approach has been popular among researchers because of its simplicity. It does not require data on the objective function, only availability of data on the practices themselves. In this regard, this approach is referred to as an "indirect" approach.

Two types of methodologies have been used to implement this approach. The first methodology relies on the idea that correlation between two practices can be interpreted as



the first evidence of complementarities. This suggests a simple test for complementarities measuring the correlation coefficient κ_C . A large value of κ_C provides evidence in favor of the existence of complementarities.

The second methodology relies on measuring correlations among error terms of equations representing the demands of practices. The adoption of the respective practices is regressed conditionally on assumed exogenous control variables, given by \mathbf{z} (we drop firm subscript for simplicity):

$$y_1^* = \alpha \mathbf{z}_1 + \varepsilon_1, \quad y_1 = 1 \quad \text{if } y_1^* > 0 \text{ and } 0 \text{ otherwise} \quad (18.29)$$

$$y_2^* = \alpha \mathbf{z}_2 + \varepsilon_2, \quad y_2 = 1 \quad \text{if } y_2^* > 0 \text{ and } 0 \text{ otherwise} \quad (18.30)$$

The error terms ($\varepsilon_1 \varepsilon_2$) are assumed to be normally distributed with zero mean ($E[\varepsilon_1] = E[\varepsilon_2] = 0$), variances equal to 1 ($Var[\varepsilon_1] = Var[\varepsilon_2] = 1$), and the covariance equal to ρ ($Cov[y_1, y_2] = \rho$).

Equations (18.29) and (18.30) can be jointly estimated by using a bivariate probit approach. In this case, a statistically significant covariance coefficient between the error terms of the regressions would imply a complementary relationship.

The adoption approach can be implemented to test complementarity among three or more practices by using a multivariate framework (Arora and Gambardella, 1990; Schmiedeberg, 2008).

Productivity approach This approach starts out with a performance equation (for example, a production function). In this case, testing for complementarities relies on regressing a measure of firm performance on interactions (i.e., combinations) of the potential complements. In practice, the implementation of this approach differs depending on whether we use dichotomous (discrete) practices or continuously measured practices.

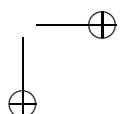
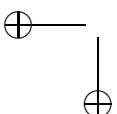
For dichotomous practices the analysis of complementarity builds on the concept of supermodularity introduced by Topkis (1978). In this case, to test the complementarity hypothesis, we need to derive an inequality restriction as implied by the theory of supermodularity and test whether this restriction is accepted by the data. For a two-dimensional function $f(y_1, y_2)$, where $y_1 = \{0, 1\}$ and $y_2 = \{0, 1\}$, practices y_1 and y_2 are (strictly) complementary if:

$$f(1, 1) - f(0, 1) > f(1, 0) - f(0, 0). \quad (18.31)$$

Equation (18.31) means that the difference in the performance function, f , that arises from starting to implement one practice (for example, from changing y_1 from 0 to 1) is greater if the other practice is also implemented (for example, greater if $y_2 = 1$ than if $y_2 = 0$).

To implement this approach empirically, the performance function is typically estimated in a multivariate regression framework as a function of four mutually exclusive combinations of the practices of interest and other exogenous factors that may affect performance (we drop firm subscript for simplicity):

$$f(y_1, y_2, \mathbf{z}) = \theta_{00}(1 - y_1)(1 - y_2) + \theta_{10}y_1(1 - y_2) + \theta_{01}(1 - y_1)y_2 + \theta_{11}y_1y_2 + \mathbf{z}\theta\mathbf{z} + \varepsilon, \quad (18.32)$$



where \mathbf{z} is a vector of exogenous variables and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The test statistic for complementarity between y_1 and y_2 now corresponds to:

$$\kappa_P \equiv \theta_{11} - \theta_{01} - \theta_{10} + \theta_{00} \quad (18.33)$$

A value of κ_P significantly greater than zero indicates that we can reject the null hypothesis of no complementarity.

The productivity approach using dichotomous practices can be generalized to the case of three or more practices. However, no simple testing procedures are available since, in this case, testing for complementarity involves adopting a multiple inequality restrictions framework. Mohnen and Röller (2005) derive the set of inequality constraints that need to be satisfied for the case of four practices. These authors proceed by testing each pair of practices separately, which implies jointly testing four inequality constraints by using a distance or Wald test proposed by Kodde and Palm (1986). Leiponen (2005) and Belderbos, Carree and Lokshin (2006) use a similar specification. Aral, Brynjolfsson and Wu (2012) and Tambe, Hitt and Brynjolfsson (2012) use a graphical framework (the *Cube View*) to understand the complementarities among three practices, where each axis represents one of the practices and there are eight potential combinations of practices. These authors propose four tests for complementarities: three specific tests of pairwise complementarities and a full test of three-way complementarities (the *System Test*). The *System Test* determines whether the gains from implementing the full system of practices are greater than the sum of gains from adopting any one of the three practices in isolation.

The productivity approach can also be implemented when practices y_1 and y_2 are measured as continuous variables. Complementarity between two continuous variables means that the incremental effect of one variable on the performance function increases conditionally on increasing the other variable. In this case, the test for complementarity relies on using a cross-term specification of the performance function. Now the performance function is estimated in a multivariate regression framework as a function of the continuous versions of the practices of interest along with an interaction term between these practices and other exogenous factors (we drop firm subscript for simplicity):

$$f(y_1, y_2, \mathbf{z}) = \alpha_1 y_1 + \alpha_2 y_2 + \alpha_{12} y_1 y_2 + \mathbf{z}\boldsymbol{\alpha}_Z + \varepsilon \quad (18.34)$$

where \mathbf{z} is a vector of exogenous variables and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The cross-derivative $\frac{\partial^2 f}{\partial y_1 \partial y_2}$ is equal to α_{12} . This implies that there is evidence for complementarity between practices y_1 and y_2 if α_{12} is significantly greater than zero.

Again, this approach can be generalized to the case of three or more practices measured as continuous variables. As in the case of dichotomous practices, difficulties arise from the need to test multiple inequality restrictions. The first studies are focused on estimating all pairwise interaction effects in one equation (see, for example, Caroli and Van Reenen, 2001), but do not consider the effect of additional cross-terms (for example, a three-way term in the case of three practices). Recent empirical studies deal with this issue. Aral et al. (2012) and Tambe et al. (2012) are also concerned with complementarities between three practices measured as continuous variables. These authors point out that a positive coefficient of the three-way term is not sufficient for complementarity. In this case, to establish the conditions for complementarity, it is necessary to evaluate the terms and cross-terms over the sample

range for each practice. Carree et al. (2011) introduce an alternative test for complementarity in the general case of n practices. These authors focus on explaining three and four practices, and claim that their procedure is also applicable to the case of dichotomous practices. Carree et al. (2011) propose a separate induced test for complementarity. In this case, the separate induced procedure accepts the null hypothesis of complementarity (which is a combined hypothesis) if and only if the separate hypotheses are all accepted. These authors derive the separate hypotheses to be held in the context of a linear regression (the number of separate hypotheses depends on the number of practices, n).

Unobserved heterogeneity Up to this point, we have omitted the discussion of the problems created by unobserved heterogeneity. Firms' practices are typically endogenous decisions, and it is precisely (unobserved) firm heterogeneity in the determinants of firms' practices that is one of the main problems in identifying complementarities. The existence of firms' unobserved heterogeneity can bias the tests for complementarities based on both the adoption and productivity approaches we described above. For example, a positive correlation between two practices or residuals cannot serve as a definite test for complementarity, as it might be the result of unobserved heterogeneity. Athey and Stern (1998) carefully discuss the problem of unobserved heterogeneity in the context of cross-section data, while Brynjolfsson and Milgrom (2013) present a recent review of the principal approaches to mitigating the effects of unobserved heterogeneity.

Athey and Stern (1998) propose using a system of equations approach, estimating the demand and the performance equations simultaneously. Kretschmer, Miravete and Pernás (2012) is the first example that implements this approach. These authors adopt the setup of Miravete and Pernás (2006) to distinguish the complementarity and correlation caused by unobserved heterogeneity. However, integrating the adoption approach and the productivity approach in a single estimation procedure is challenging. In the context of cross-section data, a two-step regression can also be estimated by using instrumental variables (Brynjolfsson and Milgrom, 2013). Moreover, if panel data are available, we can look to panel data techniques for tools to control for unobserved heterogeneity. For example, if we assume that unobserved heterogeneity does not change over time, we can control for it by including fixed effects or taking first differences (Brynjolfsson and Milgrom, 2013). Panel data also allow us to control for the simultaneity in the choices of output and inputs (Leiponen, 2005).

As pointed out by Brynjolfsson and Milgrom (2013), there are alternative approaches for mitigating the effects of unobserved heterogeneity that are not based on econometric techniques. A leading example is the use of homogeneous populations. The idea behind this approach is to eliminate as much unobserved heterogeneity as possible by identifying a narrow performance function that can be modeled empirically. In practice, this approach is typically implemented by limiting the analysis to firms in a narrow industry. The rationale is that firms in the same industry are expected to be similar in terms of the performance function. The obvious drawback of this approach is that the results may not be generalizable to other contexts.

To sum up, the study of complementarity is challenging. There are two main approaches for testing the existence of complementarities, although in practice, both tests are often useful. Moreover, unobserved heterogeneity can seriously affect both tests. One empirical strategy followed in some studies is to use both approaches to present as much evidence consistent

with the complementarity hypothesis as possible (Bresnahan, Brynjolfsson and Hitt, 2002; Cassiman and Veugelers, 2006; Ichniowski, Shaw and Prennushi, 1997). This evidence, considered as a whole, may strongly suggest the existence of such complementarity.

4.2.2 Empirical studies of complementarity in R&D activities

In what follows, we present examples of empirical studies focused on identifying complementarity relationships in R&D activities. This revision does not aim to be comprehensive, and we restrict our attention to those studies that we think are the most relevant to understanding the scope of this literature. Moreover, we are interested in the evidence based on the approaches we described above.

Before starting, it is important to notice that we focus on complementarity between practices when at least one of these practices is related to R&D. Inevitably, our review is limited, since complementarity has been empirically explored in many fields related to industrial organization, economics of innovation and strategy literature. Brynjolfsson and Milgrom (2013) and Ennen and Richter (2010) present comprehensive reviews of this broad literature. For example, empirical studies of complementarity include the relationship between new organizational practices and the use of ICTs (see Brynjolfsson and Hitt, 2000, for a review of this literature), human resource management practices (see, for example, Ichniowski et al., 1997, and Ichniowski and Shaw, 2003, for a review of this literature), different types of innovations (Miravete and Pernías, 2006; Martínez-Ros and Labeaga, 2009; Ballot et al., 2015), and obstacles to innovation (Mohnen and Rosa, 2002; Galia and Legros, 2004; Mohnen and Röller, 2005).

Researchers in the R&D literature have focused mainly on the analysis of complementarity between internal R&D and different types of external technology sourcing. Other examples of studies of complementarity in R&D activities are the relationship between partners in R&D cooperation, R&D cooperation and other practices, R&D and ICT investments, domestic and foreign R&D, R&D and human capital, and components of R&D.

Complementarity between internal R&D and external technology sourcing The view of in-house R&D as a major factor in explaining technological innovation and productivity growth is widespread. However, firms typically combine this internal activity with external sources, including, among others, R&D contracting, R&D licensing, and hiring personnel from competing firms. As we explained before, the joint adoption of such internal and external activities suggests that these activities are complementary, and most of the empirical evidence supports this result. Literature has proposed a number of reasons for complementarity between internal R&D and external technology sources. These factors are related mainly to the concepts of abortive capacity and economies of scope (see Schmiedeberg, 2008, for a detailed discussion).

One of the more detailed analyses of complementarity between internal R&D and external technology sources comes from Cassiman and Veugelers (2006). They examine 269 Belgian firms from the Community Innovation Survey (CIS)³⁸ conducted in 1993 and characterize the firm's innovation practices (internal R&D and external technology acquisition) using dichotomous variables. From a methodological point of view, Cassiman and Veugelers (2006)

³⁸ CIS data are widely used in the empirical literature (see Mairesse and Mohnen, 2010, for a review). This is also the data source used in many of the studies we review in this section.

present evidence using both the adoption and the productivity approaches. The performance measure they use is the percentage of sales due to new or substantially improved products. Moreover, they propose a two-step procedure for correcting for unobserved heterogeneity. In the first step (i.e., the adoption approach), the predicted values of the innovation practices are calculated. In the second step (i.e., the productivity approach), the predicted values of the first-step regressions are used as instruments. However, they are aware of the difficulty of finding perfectly exogenous instruments. Their two main conclusions are that (i) their results are consistent with the existence of complementarity between firms' internal R&D and external technology acquisition; and (ii) the degree of complementarity is context specific and depends on the use of "basic" R&D.

Schmiedeberg (2008) runs a similar analysis using German CIS data. However, there are two main differences with Cassiman and Veugelers (2006). First, this author does not use a two-step procedure combining the adoption and the productivity approaches. Second, this author tests for complementarity between internal R&D and externally contracted R&D, but also between internal R&D and R&D cooperation. The results of this study are twofold: (i) there is evidence of complementarity between internal R&D and R&D cooperation; but (ii) the evidence for complementarity between internal R&D and externally contracted R&D is rather weak.

Lokshin, Belderbos and Carree (2008) is a representative application of the productivity approach using continuous variables. Using a panel of Dutch manufacturing firms, the authors test for complementarity between internal and external R&D in determining labor productivity. Panel data estimation techniques allow them to control for unobserved heterogeneity and potential endogeneities. As its main result, this paper finds evidence of complementarity between internal and external R&D, but only when allowing for diseconomies of scale in internal and external R&D. In this case, the positive impact of external R&D is conditional upon a sufficient level of internal R&D. More recently, Hagedoorn and Wang (2012) find similar results in their study of pharmaceutical firms. They find evidence of complementarity (substitutability) between internal and external R&D if the level of internal R&D investment is high (low).

In a recent study, and going beyond the existence of complementarity, Ceccagnoli, Higgins and Palermo (2014) focus on the conditions under which internal R&D and external technology sources are complements (i.e., the study of "complementarity drivers"). As pointed out by these authors, there is little research evidence on this issue (Cassiman and Veugelers, 2006 is one exception). Similar to Lokshin et al. (2008), Ceccagnoli et al. (2014) apply panel data estimation techniques to a panel of 94 pharmaceutical firms to study whether internal R&D investments and in-licensing expenditures are complementary in determining the firm's product pipeline. Using the whole sample of firms, they find that internal R&D and in-licensing expenditures are neither complements nor substitutes. However, and more interestingly, this result does not hold when analyzing the impact of three potential drivers of complementarity: (i) absorptive capacity; (ii) economies of scope; and (iii) licensing experience. The idea is to separately analyze subsamples of firms with low and high levels of the examined drivers. Doing this, they find a complementary relationship between internal R&D and in-licensing for firms that have a larger value (to 25 percent) of two of the drivers (economies of scope and licensing experience).

The literature we have reviewed so far uses data from developed countries. However, a number of empirical studies have also analyzed the complementarity between internal R&D

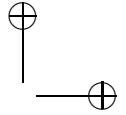
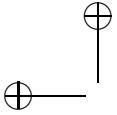
and external technology sourcing in developing countries (see Hou and Mohnen, 2013, for a review of this literature). As a leading example of this literature, Hu, Jefferson and Qian (2005) analyze the complementarity among three continuous variables (internal R&D and expenditures on disembodied technology from foreign and domestic providers) in determining firm productivity. One limitation of this study is that it estimates three pairwise interaction effects in one equation (a production function), but does not consider the effect of a three-way term. Using a panel of approximately 10 000 Chinese firms, these authors find evidence of complementary relationships between internal R&D and both the domestic and foreign technology transfer variables.

Other topics in the literature of R&D complementarity The determinants and the effects of R&D cooperation with different partners have been widely investigated. However, little evidence exists on the complementarity effect between R&D cooperation with different partners. Using Dutch CIS data, Belderbos, Carree and Lokshin (2006) are the first to explore the complementarity effects of four types of R&D cooperation on labor productivity.³⁹ This is an example of the application of the productivity approach using dichotomous variables (in this case, four dummies for R&D cooperation with competitors, customers, suppliers, and universities and research institutes). Their empirical results on complementarities are mixed. On the one hand, they find evidence of complementarity between R&D cooperation with competitors and with customers, and between R&D cooperation with customers and with universities and research institutes. On the other hand, they find evidence, especially for small firms, of substitutability between three pairs of practices (supplier cooperation combined with either competitor or university and research institute cooperation, and competitor cooperation combined with university and research institute cooperation).

A number of studies have also analyzed R&D cooperation, but in terms of its potential complementarity with other practices (in this line, we have already cited the paper by Schmiedeberg, 2008). Using Finnish CIS data, Leiponen (2005) finds a significant complementarity effect between R&D cooperation and technical skills (measured by educational levels and fields of employees) in determining a firm's profits. More recently, Harhoff, Müller and Van Reenen (2014) analyze technology-sourcing activities of German firms in the USA. They find a complementarity effect between R&D cooperation with suppliers located in Germany (USA) and the R&D stock in Germany (USA) at the industry level (R&D stocks are used as a proxy for the local knowledge pools). Finally, Arvanitis et al. (2015) use Swiss and Dutch CIS data to analyze the complementarity between external R&D and R&D cooperation in the presence of internal R&D. These authors find little evidence supporting this hypothesis.

Other examples of studies of complementarity in R&D activities are the relationship between R&D and ICT investments (Hall, Lott and Mairesse, 2013), domestic and foreign R&D (Belderbos, Lokshin and Sadowski, 2015), R&D and human capital (Ballot, Fakhfakh and Taymaz, 2001), and components of R&D (Barge-Gil and López, 2013).

³⁹ Veugelers and Cassiman (2005) analyze the relationship between R&D cooperation between firms (customers and/or suppliers) and universities. However, as they pointed out, they do not test for complementarity as such. In terms of the empirical approaches reviewed above, they only present pairwise correlations between R&D cooperation with firms and R&D cooperation with universities.



5 CONCLUSIONS

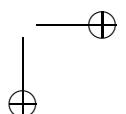
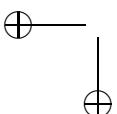
The objective of this chapter has been to provide a general view of three of the main topics in current empirical literature about firms' R&D: the determinants of R&D investments, the link between R&D, innovation and productivity, and the studies trying to open up and examine the contents of the black box of R&D. Regarding research questions, our selection of subjects indicates that classical topics such as the role of size and market power in determining R&D or the effect of R&D on productivity still receive considerable attention. However, nowadays there is a tendency to focus on other issues: the role played by public funding as an R&D determinant, the channels (for example, new products and processes) through which R&D influences productivity, the importance of R&D composition and the complementarities of internal R&D with other activities, such as external R&D, cooperation, ICT investments or human capital. Other topics that fall beyond the scope of the present chapter are also receiving increasing attention: the relationship between R&D and employment, the causes and effects of international R&D sourcing or, in association with the recent economic crisis, the response of R&D investment to economic cycles.

Our review also points out that there is a tendency towards a higher use of structural theoretical models to support empirical analysis. These models have made it possible: (i) to distinguish between the determinants of R&D-extensive margin (the decision to undertake R&D activities) and those of R&D-intensive margin (the magnitude or intensity of R&D expenditures); (ii) to reveal the problems of uncertainty and incomplete information that are usually present in R&D markets; (iii) to link additionality effects of public support of private R&D (when they exist) with social welfare; (iv) to take into account the persistence and the dynamic nature of R&D decisions and their effects (as long as there is a time lag between R&D investment and output, R&D is unlikely to have a one-time impact and the gains in profits derived from R&D are surrounded by uncertainty); (v) to quantify the effect of changes in the firm's environment on its decision to invest in R&D and long-run profitability.

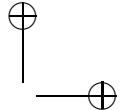
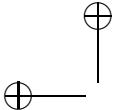
An important question remains open: where should empirical research on business R&D go now? There are at least two related areas in which this research may be extended: methodology and data availability. Regarding the first, although there has been significant progress in the use of structural modeling, this is still an emerging line of work. A second methodological issue concerns the exploitation and design of controlled field experiments. The experimental approach may complement what we learn from more sophisticated modeling. On the data side, one interesting opportunity is the use of longer panels that would allow for a more appropriate analysis of the relationship between R&D and the dynamics of entry and exiting of firms, and also of the persistence in firms' behavior differences. The availability of more quantitative data would also improve qualitative analyses that are frequently conditioned by the categorical character of innovation outputs in public databases.

REFERENCES

- Acemoglu, D. and J. Linn (2004): "Market size in innovation: theory and evidence from the pharmaceutical industry", *The Quarterly Journal of Economics* 119(3), 1049–1090.
Aerts, K. and S. Thorwarth (2008): "Additionality effects of public R&D funding: 'R' versus 'D'", *FBE Research Report MSI_0811*.
Aghion, P. and P. Howitt (1992): "A model of growth through creative destruction", *Econometrica* 60, 323–351.

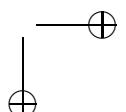
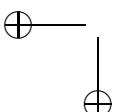


- Aghion, P., N. Bloom and R. Blundell et al. (2005): "Competition and innovation: an inverted-U relationship", *The Quarterly Journal of Economics* 120(2), 701–728.
- Aguirregabiria, V. and P. Mira (2007): "Sequential estimation of dynamic discrete games", *Econometrica* 75(1), 1–53.
- Anderson, W.H.L. (1967): "Business fixed investment: a marriage of fact and fancy", in R. Ferber (ed.), *Determinants of Investment Behavior*. New York: Columbia University Press, 413–425.
- Aral, S., E. Brynjolfsson and L. Wu (2012): "Three-way complementarities: performance pay, human resource analytics, and information technology", *Management Science* 58(5), 913–931.
- Arora, A. and M. Ceccagnoli (2006): "Patent protection, complementary assets and firms' incentives for technology licensing", *Management Science* 52(2), 293–308.
- Arora, A. and A. Gambardella (1990): "Complementarity and external linkages: the strategies of the large firms in biotechnology", *The Journal of Industrial Economics* 38(4), 361–379.
- Arora, A., M. Ceccagnoli and W.M. Cohen (2008): "R&D and the patent premium", *International Journal of Industrial Organization* 26(5), 1153–1179.
- Arqué-Castells, P. (2013): "Persistence in R&D performance and its implications for the granting of subsidies", *Review of Industrial Organization* 43, 193–220.
- Arqué-Castells, P. and P. Mohnen (2015): "Sunk costs, extensive R&D subsidies and permanent inducement effects", *Journal of Industrial Economics* 63(3), 458–494.
- Arrow, K. (1962): "Economic welfare and the allocation of resources for invention", in NBER (ed.), *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Princeton, NJ: Princeton University Press, 609–626.
- Arvanitis, S., B. Lokshin, P. Mohnen and M. Woerter (2015): "Impact of external knowledge acquisition strategies on innovation: a comparative study based on Dutch and Swiss panel data", *Review of Industrial Organization* 46(4), 359–382.
- Asheim, B.T. and L. Coenen (2005): "Knowledge bases and regional innovation systems: comparing Nordic clusters", *Research policy* 34(8), 1173–1190.
- Athey, S. and S. Stern (1998): "An empirical framework for testing theories about complementarity in organizational design", *NBER Working Paper* 6600.
- Aw, B.Y., M.J. Roberts and D.J. Xu (2011): "R&D investment, exporting, and productivity dynamics", *American Economic Review* 101(4), 1312–1344.
- Ballot, G., F. Fakhfakh and E. Taymaz (2001): "Firms' human capital, R&D and performance: a study on French and Swedish firms", *Labour Economics* 8(4), 443–462.
- Ballot, G., F. Fakhfakh, F. Galia and A. Salter (2015): "The fateful triangle: Complementarities in performance between product, process and organizational innovation in France and the UK", *Research Policy* 44(1), 217–232.
- Barge-Gil, A. and A. López (2013): "The complementarity effect of research and development on firm productivity", *Applied Economics Letters* 20(15), 1426–1430.
- Barge-Gil, A. and A. López (2014): "R&D determinants: accounting for the differences between research and development", *Research Policy* 43(9), 1634–1648.
- Barge-Gil, A. and A. López (2015): "R versus D: estimating the differentiated effect of research and development on innovation results", *Industrial and Corporate Change* 24(1), 93–129.
- Bartelsman, E. and E.J. Dhrymes (1998): "Productivity dynamics: U.S. manufacturing plants, 1972–1986", *Journal of Productivity Analysis* 9, 5–34.
- Becker, B. (2013): "The determinants of R&D investment: a survey of the empirical research", *Discussion Paper Series* 2013.09, Department of Economics, Loughborough University, UK.
- Belderbos, R., M. Carree and B. Lokshin (2006): "Complementarity in R&D cooperation strategies", *Review of Industrial Organization* 28(4), 401–426.
- Belderbos, R., B. Lokshin and B. Sadowski (2015): "The returns to foreign R&D", *Journal of International Business Studies* 46, 491–504.
- Beneito, P., P. Coscollá-Girona, M.E. Rochina-Barrachina and A. Sanchis (2015): "Competitive pressure and innovation at the firm level", *The Journal of Industrial Economics* 63(3), 422–457.
- Bessen, J. and E. Maskin (2009): "Sequential innovation, patents, and imitation", *The RAND Journal of Economics* 40(4), 611–635.
- Bloch, C. (2005): "R&D investment and internal finance: the cash flow effect", *Economics of Innovation and New Technology* 14(3), 213–223.
- Blundell, R., R. Griffith and J. van Reenen (1999): "Market share, market value and innovation in a panel of British manufacturing firms", *The Review of Economic Studies* 66(3), 529–554.
- Bond, S., D. Harhoff and J. van Reenen (2005): "Investment, R&D and financial constraints in Britain and Germany", *Annales d'Economie et de Statistique* 79/80, 433–460.
- Borisova, G. and J.R. Brown (2013): "R&D sensitivity to asset sale proceeds: new evidence on financing constraints and intangible investment", *Journal of Banking and Finance* 37(1), 159–173.
- Bresnahan, T.F., E. Brynjolfsson and L.M. Hitt (2002): "Information technology, workplace organization, and the demand for skilled labor: firm level evidence", *Quarterly Journal of Economics* 117, 339–376.



512 *Handbook of game theory and industrial organization: applications*

- Brown, J.R., S.M. Fazzari and B.C. Petersen (2009): "Financing innovation and growth: cash flow, external equity, and the 1990s R&D boom", *The Journal of Finance* 64(1), 151–185.
- Brynjolfsson, E. and L.M. Hitt (2000): "Beyond computation: information technology, organizational transformation and business performance", *Journal of Economic Perspectives* 14(4), 23–48.
- Brynjolfsson, E. and P. Milgrom (2013): "Complementarity in organizations", in R. Gibbons and J. Roberts (eds), *Handbook of Organizational Economics*. Princeton, NJ: Princeton University Press.
- Cabral, L. (1994): "Bias in market R&D portfolios", *International Journal of Industrial Organization* 12(4), 533–547.
- Cabral, L. (2003): "R&D competition when firms choose variance", *Journal of Economics and Management Strategy* 12(1), 139–150.
- Caroli, E. and J. van Reenen (2001): "Skill-biased organizational change: evidence from a panel of British and French establishments", *Quarterly Journal of Economics* 116(4), 1449–1492.
- Carree, M., B. Lokshin and R. Belderbos (2011): "A note on testing for complementarity and substitutability in the case of multiple practices", *Journal of productivity Analysis* 35, 263–269.
- Cassiman, B. and R. Veugelers (2006): "In search of complementarity in innovation strategy: internal R&D and external knowledge acquisition", *Management Science* 52(1), 68–82.
- Ceccagnoli, M., M.J. Higgins and V. Palermo (2014): "Behind the scenes: sources of complementarity in R&D", *Journal of Economics and Management Strategy* 23(1), 125–148.
- Cefis, E. (2003): "Is there persistence in innovative activities?" *International Journal of Industrial Organization* 21, 489–515.
- Cefis, E. and M. Ciccarelli (2005): "Profit differentials and innovation", *Economics of Innovation and New Technology* 14, 43–61.
- Cefis, E. and L. Orsenigo (2001): "The persistence of innovative activities. A cross-countries and cross-sectors comparative analysis", *Research Policy* 30, 1139–1158.
- Chiesa, V. (2001): *R&D Strategy and Organisation: Managing Technical Change in Dynamic Contexts, Series on Technology Management: Volume 5*. Singapore: World Scientific Publishing Company.
- Clausen, T.H. (2009): "Do subsidies have positive impacts on R&D and innovation activities at the firm level?" *Structural Change and Economic Dynamics* 20(4), 239–253.
- Cohen, W.M. (2010): "Fifty years of empirical studies of innovative activity and performance", in B.H. Hall and N. Rosenberg (eds), *Handbook of the Economics of Innovation, Vol. 1*. Amsterdam: Elsevier, 129–213.
- Cohen, W.M. and S. Klepper (1996): "A reprise of size and R&D", *Economic Journal* 106, 925–951.
- Cohen, W.M. and R.C. Levin (1989): "Empirical studies of innovation and market structure", in M. Armstrong and R. Porter (eds), *Handbook of Industrial Organization*, Vol. 2. Amsterdam: North-Holland, 1059–1107.
- Cohen, W.M., R.C. Levin and D.C. Mowery (1987): "Firm size and R&D intensity: a re-examination", *Journal of Industrial Economics* 35(4), 543–565.
- Collard-Wexler, A. (2013): "Demand fluctuations in the ready-mix concrete industry", *Econometrica* 81(3), 1003–1037.
- Crépon, B. and E. Duguet (1997): "Estimating the knowledge production function from patent numbers: GMM on count panel data with multiplicative errors", *Journal of Applied Econometrics* 12(3), 243–263.
- Crépon, B., E. Duguet and J. Mairesse (1998): "Research, innovation and productivity: an econometric analysis at the firm level", *Economics of Innovation and New Technology* 7, 115–158.
- Czarnitzki, D. and H. Hottenrott (2011): "Financial constraints: routine versus cutting edge R&D investment", *Journal of Economics and Management Strategy* 20(1), 121–157.
- Czarnitzki, D. and S. Thorwarth (2012): "Productivity effects of basic research in low-tech and high-tech industries", *Research Policy* 41(9), 1555–1564.
- Czarnitzki, D. and A. Toole (2011): "Patent protection, market uncertainty, and R&D investment", *The Review of Economics and Statistics* 93(1), 147–159.
- Czarnitzki, D., H. Hottenrott and S. Thorwarth (2011): "Industrial research versus development investment: the implications of financial constraints", *Cambridge Journal of Economics* 35(3), 527–544.
- Czarnitzki, D., K. Kraft and S. Thorwarth (2009): "The knowledge production of 'R' and 'D'", *Economics Letters* 105(1), 141–143.
- Das, S., R. Mark and J.R. Tybout (2007): "Market entry costs, producer heterogeneity, and export dynamics", *Econometrica* 75(3), 837–873.
- David, P.A., B.H. Hall and A.A. Toole (2000): "Is public R&D a complement or substitute for private R&D? A review of the econometric evidence", *Research Policy* 29, 497–529.
- Deschryvere, M. (2014): "R&D, firm growth and the role of innovation persistence: an analysis of Finnish SMEs and large firms", *Small Business Economics* 43, 767–785.
- Dobbelaere, S. and J. Mairesse (2010): "Micro-evidence on rent sharing from different perspectives", *NBER Working Paper* 16220.
- Dobbelaere, S. and J. Mairesse (2013): "Panel data estimates of the production function and product and labor market imperfections", *Journal of Applied Econometrics* 28, 1–46.

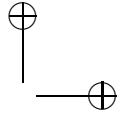
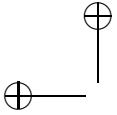


- Doraszelski, U. and J. Jaumandreu (2013): "R&D and productivity: estimating endogenous productivity", *Review of Economic Studies* 80, 1338–1383.
- Dorfman, R. and Steiner, P.O. (1954): "Optimal advertising and optimal quality", *American Economic Review* 44, 826–836.
- Duguet, E. and C. Lelarge (2012): "Does patenting increase the private incentives to innovate? A microeconometric analysis", *Annales d'Economie et de Statistique* 107/108, 201–238.
- Duguet, E. and S. Monjon (2004): "Is innovation persistent at the firm level? An econometric examination comparing the propensity score and regression methods", *Cahiers de la Maison des Sciences Economiques* v04075, Université Panthéon-Sorbonne (Paris 1).
- Encaoua, D., B.H. Hall, F. Laisney and J. Mairesse (eds) (2013): *The Economics and Econometrics of Innovation*. New York: Springer Science and Business Media.
- Ennen, E. and A. Richter (2010): "The whole is more than the sum of its parts – or is it? A review of the empirical literature on complementarities in organizations", *Journal of Management* 36(1), 207–233.
- Fariñas, J.C. and S. Ruano (2005): "Firm productivity, heterogeneity, sunk costs and market selection", *International Journal of Industrial Organization* 23, 505–534.
- Flaig, G. and M. Stadler (1994): "Success breeds success: the dynamics of the innovation process," *Empirical Economics* 19, 55–68.
- Galia, F. and D. Legros (2004): "Complementarities between obstacles to innovation: evidence from France", *Research Policy* 33, 1185–1199.
- Geraski, P.A. (1990): "Innovation, technological opportunity, and market structure", *Oxford Economic Papers* 42(3), 586–602.
- Geraski, P.A. and C.F. Walters (1995): "Innovative activities over the business cycle", *The Economic Journal* 105, 916–928.
- Geraski, P.A., J. van Reenen and C.F. Walters (1997): "How persistently do firms innovate?" *Research Policy* 26, 33–48.
- González, X., J. Jaumandreu and C. Pazó (2005): "Barriers to innovation and subsidy effectiveness", *The RAND Journal of Economics* 36(4), 930–949.
- Graevenitz, G., S. Wagner and D. Harhoff (2013): "Incidence and growth of patent thickets: the impact of technological opportunities and complexity", *The Journal of Industrial Economics* 61(3), 521–563.
- Griliches, Z. (1979): "Issues in assessing the contribution of research and development to productivity growth", *Bell Journal of Economics* 10, 92–116.
- Griliches, Z. (1986): "Productivity, R&D, and basic research at the firm level in the 1970's", *The American Economic Review* 76(1), 141–154.
- Griliches, Z. and F. Lichtenberg (1984): "R&D and productivity growth at the industry level: is there still a relationship?" in Z. Griliches (ed.), *R&D, Patents and Productivity*. Chicago, IL: University of Chicago Press, 465–502.
- Grossman, G. and E. Helpman (1991): *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.
- Hagedoorn, J. and N. Wang (2012): "Is there complementarity or substitutability between internal and external R&D strategies?" *Research Policy* 41(6), 1072–1083.
- Hall, B.H. (2002): "The financing of research and development", *Oxford Review of Economic Policy* 18(1), 35–51.
- Hall, B.H. (2011): "Innovation and productivity", *Nordic Economic Policy Review* 2, 167–204.
- Hall, B.H. and J. Lerner (2010): "The financing of R&D and innovation", in B.H. Hall and N. Rosenberg (eds), *Handbook of the Economics of Innovation, Vol. 1*. Amsterdam: Elsevier, 609–639.
- Hall, B.H. and J. Mairesse (1995): "Exploring the relationship between R&D and productivity in French manufacturing firms", *Journal of Econometrics* 65(1), 263–293.
- Hall, B.H. and V. Sena (2014): "Appropriability mechanisms, innovation and productivity: evidence from the UK", *NBER Working Paper* w20514.
- Hall, B.H. and R.H. 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.
- Hall, B., C. Helmers, M. Rogers and V. Sena (2014): "The choice between formal and informal intellectual property: a review", *Journal of Economic Literature* 52(2), 375–423.
- Hall, B.H., F. Lott and J. Mairesse (2013): "Evidence on the impact of R&D and ICT investments on innovation and productivity in Italian firms", *Economics of Innovation and New Technology* 22(3), 300–328.
- Hall, B.H., J. Mairesse and P. Mohnen (2010): "Measuring the returns to R&D", in B.H. Hall and N. Rosenberg (eds), *Handbook of the Economics of Innovation, Vol. 2*, Amsterdam: Elsevier, 1034–1076.
- Hall, R.E. (1988): "The relationship between price and marginal cost in U.S. industry", *Journal of Political Economy* 96, 921–47.
- Harhoff, D., E. Müller and J. van Reenen (2014): "What are the channels for technology sourcing? Panel data evidence from German companies", *Journal of Economics and Management Strategy* 23(1), 204–224.
- Harris, M.N., M. Rogers and A. Siouclis (2003): "Modeling firm innovation using panel probit estimators", *Applied Economics Letters* 10(11), 683–686.

514 *Handbook of game theory and industrial organization: applications*

- Heckman, J. (1976): "The common structure of statistical models of truncation. Sample selection and limited variables and a simple estimator for such models", *Annals of Economic and Social Measurement* 5, 475–492.
- Heckman, J. (1979): "Sample selection bias as specification error", *Econometrica* 47, 153–161.
- Heckman J.J. (1981): "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process", in C. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press, 179–195.
- Henderson, R. and I. Cockburn (1996): "Scale, scope, and spillovers: the determinants of research productivity in drug discovery", *The RAND Journal of Economics* 27(1), 32–59.
- Hou, J. and P. Mohnen (2013): "Complementarity between in-house R&D and technology purchasing: evidence from Chinese manufacturing firms", *Oxford Development Studies* 41(3), 343–371.
- Howe, J.D. and D.G. McFetridge (1976): "The determinants of R&D expenditures", *Canadian Journal of Economics* 9, 57–71.
- Hu, A., G. Jefferson and J. Qian (2005): "R&D and technology transfer: firm-level evidence from Chinese industry", *Review of Economics and Statistics* 87(4), 780–786.
- Huergo, E. and L. Moreno (2011): "Does history matter for the relationship between R&D, innovation and productivity?" *Industrial and Corporate Change*, 20(5), 1335–1368.
- Ichniowski, C. and K. Shaw (2003): "Beyond incentive pay: Insiders' estimates of the value of complementary human resource management practices", *Journal of Economic Perspectives* 17(1), 155–78.
- Ichniowski, C., K. Shaw and G. Prennushi (1997): "The effects of human resource management practices on productivity: a study of steel finishing lines", *American Economic Review* 87(3), 291–313.
- Kamien, M.I. and N.L. Schwartz (1970): "Market structure, elasticity of demand, and incentive to invent", *Journal of Law and Economics* 13, 241–252.
- Kamien, M.I. and N.L. Schwartz (1976): "On the degree of rivalry for maximum innovative activity", *Quarterly Journal of Economics* 90, 245–260.
- Kamien, M.I. and N.L. Schwartz (1978): "Self-financing of an R&D project", *American Economic Review* 68, 252–261.
- Kamien, M.I. and Schwartz, N.L. (1982): *Market Structure and Innovation*. Cambridge, UK: Cambridge University Press.
- Klette, J. (1996): "R&D, scope economies, and plant performance", *RAND Journal of Economics*, 27, 502–522.
- Klette, J. and Z. Griliches (1996): "The inconsistency of common scale estimators when output prices are unobserved and endogenous", *Journal of Applied Econometrics* 11, 343–361.
- Klevorick, A.K., R.C. Levin, R.R. Nelson and S.G. Winter (1995): "On the sources and significance of interindustry differences in technological opportunities", *Research Policy* 24(2), 185–205.
- Klomp, L. and G. van Leeuwen (2001): "Linking innovation and firm performance: a new approach", *International Journal of the Economics of Business* 8, 343–364.
- Kodde, D.A. and F.C. Palm (1986): "Wald criteria for jointly testing equality and inequality restrictions", *Econometrica* 54(5), 1243–1248.
- Kretschmer, T., E.J. Miravete and J.C. Pernias (2012): "Competitive pressure and the adoption of complementary innovations", *American Economic Review* 102(4), 1540–1570.
- Kwon, I. (2010): "R&D portfolio and market structure", *Journal of Industrial Economics* 120, 313–323.
- Lach, S. and M. Schankerman (1989): "Dynamics of R&D and investment in the scientific sector", *The Journal of Political Economy* 97(4), 880–904.
- Leifer, R. and T. Triscari (1987): "Research versus development: differences and similarities", *IEEE Transactions on Engineering Management* 34(2), 71–78.
- Leiponen, A. (2005): "Skills and innovation", *International Journal of Industrial Organization* 23, 303–323.
- Lerner, J. (2009): "The empirical impact of intellectual property rights on innovation: puzzles and clues", *The American Economic Review* 99(2), 343–348.
- Levin, R.C., W.M. Cohen and D.C. Mowery (1985): "R&D appropriability, opportunity, and market structure: new evidence on some Schumpeterian hypotheses", *The American Economic Review* 75(2), 20–24.
- Lichtenberg, F.R. and D. Siegel (1991): "The impact of R&D Investment on productivity – new evidence", *Economic Inquiry* 29(2), 203.
- Lim, K. (2004): "The relationship between research and innovation in the semiconductor and pharmaceutical industries (1981–1997)", *Research Policy* 33, 287–321.
- Link, A.N. (1981): "Basic research and productivity increase in manufacturing: additional evidence", *American Economic Review* 71(5), 1111–1112.
- Link, A.N. (1982): "An analysis of the composition of R&D spending", *Southern Economic Journal* 49(2), 342–349.
- Link, A.N. (1985): "The changing composition of R&D", *Managerial and Decision Economics* 6(2), 125–128.
- Lokshin, B., R. Belderbos and M. Carree (2008): "The productivity effects of internal and external R&D: evidence from a dynamic panel data model", *Oxford Bulletin of Economics and Statistics* 70, 399–413.
- Mairesse, J. and J. Jaumandreu (2005): "Panel-data estimates of the production function and the revenue function: what difference does it make?", *Scandinavian Journal of Economics* 107, 651–672.

- Mairesse, J. and P. Mohnen (2010): "Using innovation surveys for econometric analysis", in B.H. Hall and N. Rosenberg (eds), *Handbook of the Economics of Innovation, Vol. 2*. Amsterdam: Elsevier, 1129–1156.
- Mairesse, J. and M. Sassenou (1991): "R&D productivity: a survey of econometric studies at the firm level", *NBER Working Papers* 3666.
- Mairesse, J., B.H. Hall and B. Mulkay (1999): "Firm-level investment in France and the United States: an exploration of what we have learned in twenty years", *Annales d'Economie et de Statistique* 55/56, 27–67.
- Malerba, F. and L. Orsenigo (1999): "Technological entry, exit and survival: an empirical analysis of patent data", *Research Policy* 28, 643–660.
- Mansfield, E. (1980): "Basic research and productivity increase in manufacturing", *The American Economic Review* 70(5), 863–873.
- Mansfield, E. (1981): "Composition of R&D expenditures: relationship to size of firm, concentration, and innovative output", *The Review of Economics and Statistics* 63(4), 610–615.
- Mañez-Castillejo, J.A., M.E. Rochina-Barrachina, A. Sanchis-Llopis and J. Sanchis-Llopis (2009): "The role of sunk costs in the decision to invest in R&D", *Journal of Industrial Economics* 57, 712–735.
- Martin, S. (2010): *Industrial Organization in Context*. Oxford: Oxford University Press.
- Martínez-Ros, E. and J.M. Labeca (2009): "Product and process innovations: persistence and complementarities", *European Management Review* 6(1), 64–75.
- Milgrom, P. and J. Roberts (1990): "The economics of modern manufacturing: technology, strategy, and organization", *American Economic Review* 80(3), 511–528.
- Milgrom, P. and J. Roberts (1995): "Complementarities and fit: strategy, structure and organizational change in manufacturing", *Journal of Accounting and Economics* 19(2–3), 179–208.
- Milgrom, P. and J. Roberts (1996): "The LeChatelier principle", *American Economic Review* 86(1), 173–179.
- Milgrom, P. and C. Shannon (1994): "Monotone comparative statics", *Econometrica* 62, 157–180.
- Milgrom, P., Y. Qian and J. Roberts (1991): "Complementarities, momentum, and the evolution of modern manufacturing", *American Economic Association Papers and Proceedings*, 85–89.
- Miravete, E.J. and J.C. Pernías (2006): "Innovation complementarity and scale of production", *Journal of Industrial Economics* 54(1), 1–29.
- Mohnen, P. and B.H. Hall (2013): "Innovation and productivity: an update", *Eurasian Business Review* 3(1), 47–65.
- Mohnen, P. and L.H. Röller (2005): "Complementarities in innovation policy", *European Economic Review* 49, 1431–1450.
- Mohnen, P. and J. Rosa (2002): "Barriers to innovation in service industries in Canada", in M. Feldman and N. Massard (eds), *Institutions and Systems in the Geography of Innovation*. Boston, MA: Kluwer Academic Publishers, 231–250.
- Needham, D. (1975): "Market structure and firms' R&D behaviour", *Journal of Industrial Economics* 23, 241–255.
- Nelson, R.R. (1959): "The simple economics of basic scientific research", *The Journal of Political Economy* 67(3), 297–306.
- Nelson, R.R. and S.G. Winter (1982): *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap Press of Harvard University Press.
- OECD/Eurostat (2005): *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data*, 3rd edition. Paris: OECD Publishing, available at <http://dx.doi.org/10.1787/9789264013100-en>.
- Olley, S.G. and A. Pakes (1996): "The dynamics of productivity in the telecommunications equipment industry", *Econometrica* 64, 1263–1297.
- Peters, B. (2009): "Persistence of innovation: stylised facts and panel data evidence", *Journal of Technology Transfer* 34(2), 226–243.
- Peters, B., M.J. Roberts, V.A. Vuong and H. Fryges (2013): "Estimating dynamic R&D demand: an analysis of costs and long-run benefits", *NBER Working Papers* w19374.
- Png, I. (2017): "Law and innovation. Evidence from state trade secret laws", *Review of Economics and Statistics*, 99(1), doi: 10.1162/REST_a_00632.
- Raymond, W., J. Mairesse, P. Mohnen and F. Palm (2015): "Dynamic models of R&D, innovation and productivity: panel data evidence for Dutch and French manufacturing", *European Economic Review* 78, 285–306.
- Raymond, W., P. Mohnen, F. Palm and S. Schim van der Loeff (2007): "The behavior of the maximum likelihood estimator of dynamic panel data sample selection models", *CESifo Working Paper* 1992.
- Raymond, W., P. Mohnen, F. Palm and S. Schim van der Loeff (2009): "Innovative sales, R&D and total innovation expenditures: panel evidence on their dynamics", *CESifo Working Paper Series* 2716.
- Raymond, W., P. Mohnen, F. Palm and S. Schim van der Loeff (2010): "Persistence of innovation in Dutch manufacturing: is it spurious?" *The Review of Economics and Statistics* 92(3), 495–504.
- Roberts, M.J. and V.A. Vuong (2013): "Empirical modeling of R&D demand in a dynamic framework", *Applied Economic Perspectives and Policy* 35(2) 185–205.
- Rogers, M. (2004): "Networks, firm size and innovation", *Small Business Economics* 22(2), 141–153.
- Rust, J. (1987): "Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher", *Econometrica* 55(5), 999–1033.



516 *Handbook of game theory and industrial organization: applications*

- Scherer, F.M. (1980): *Industrial Market Structure and Economic Performance*, 2nd edition. Chicago, IL: Rand McNally.
- Schmiedeberg, C. (2008): "Complementarities of innovation activities: an empirical analysis of the German manufacturing sector", *Research Policy* 37, 1492–1503.
- Schmookler, J. (1962): "Economic sources of inventive activity", *The Journal of Economic History* 22(1), 1–20.
- Schumpeter, J.A. (1939): *Business Cycles*. New York: McGraw-Hill.
- Schumpeter, J.A. (1942): *Capitalism, Socialism and Democracy*. New York: Harper & Row.
- Solow, R. (1957): "Technical change and the aggregate production function", *The Review of Economics and Statistics* 39(3), 312–320.
- Spence, A.M. (1975): "Monopoly, quality, and regulation", *Bell Journal of Economics* 6, 417–429.
- Spencer, B.J. and J.A. Brander (1983): "International R&D rivalry and industrial strategy", *The Review of Economic Studies* 50(4), 707–722.
- Takalo, T., T. Tanayama and O. Toivanen (2013a): "Market failures and the additionality effects of public support to private R&D: theory and empirical implications", *International Journal of Industrial Organization* 31(5), 634–642.
- Takalo, T., T. Tanayama and O. Toivanen (2013b): "Estimating the benefits of targeted R&D subsidies", *The Review of Economics and Statistics* 95(1), 255–272.
- Tambe, P., L.M. Hitt and E. Brynjolfsson (2012): "The extroverted firm: how external information practices affect innovation and productivity", *Management Science* 58(5), 843–859.
- Topkis, D.M. (1978): "Minimizing a submodular function on a lattice", *Operations Research* 26, 305–321.
- Van Beveren, I. (2012): "Total factor productivity estimation", *Journal of Economic Surveys* 26(1), 98–128.
- Van Leeuwen, G. (2002): "Linking innovation to productivity growth using two waves of the Community Innovation Survey", *OECD Science, Technology and Industry Working Papers* 2002/8.
- Van Leeuwen, G. and L. Klomp (2006): "On the contribution of innovation to multi-factor-productivity growth", *Economics of Innovation and New Technology* 15(4–5), 367–390.
- Veugelers, R. and B. Cassiman (2005): "R&D cooperation between firms and universities. Some empirical evidence from Belgian manufacturing", *International Journal of Industrial Organization* 23(5–6), 355–379.
- Vives, X. (1990): "Nash equilibrium with strategic complementarities", *Journal of Mathematical Economics* 19, 305–321.
- Wieser, R. (2005): "R&D, productivity and spillovers: empirical evidence at firm level", *Journal of Economic Surveys* 19(4), 587–621.
- Wooldridge, J. (2005): "Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity", *Journal of Applied Econometrics*, 20(1), 39–54.
- Zúñiga-Vicente, J.A., C. Alonso-Borrego, F.J. Forcadell and J.I. Galán (2014): "Assessing the effect of public subsidies on firm R&D investment: a survey", *Journal of Economic Surveys* 28(1), 36–67.

