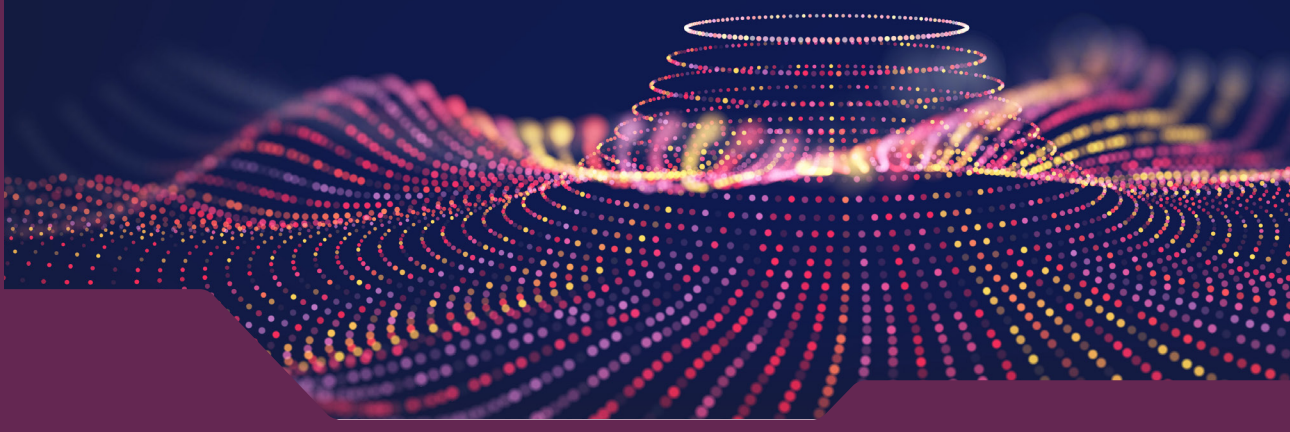
ECONOMIC ANALYSIS OF THE DIGITAL REVOLUTION

Edited by **Juan José Ganuza** and **Gerard Llobet**

FUNCAS Social and Economic Studies, 5



funcas

ECONOMIC ANALYSIS OF THE DIGITAL REVOLUTION

Juan-José Ganuza and Gerard Llobert *(editors)*

FUNCAS Social and Economic Studies, 5

FUNCAS
Madrid, Spain

Printed in Spain
Edit: FUNCAS
Caballero de Gracia, 28, 28013 - Madrid (Spain)

ISBN: 978-84-15722-95-3
ISBN: 978-84-17609-00-9
ISBN: 978-84-15722-94-6
Depósito legal: M-30350-2018
Prints: Cecabank

Index

Index

Contributors

Luis AGUIAR
Digital Economy Unit, Joint Research Center, European Commission

Paul BELLEFLAMME
Université Catholique de Louvain

Carlos BELLÓN
Universidad Pontificia Comillas (ICADE)

Adina CLAICI
Copenhagen Economics, Brussels and College of Europe, Bruges

Francesco DECAROLIS
Bocconi University

Juan-José GANUZA
Universitat Pompeu Fabra, Barcelona GSE and Funcas

Maris GOLDMANIS
University of London

Guillaume HAERINGER
Baruch College (New York)

Hanna HALABURDA
Bank of Canada and New York University

Stephen HANSEN
University of Oxford and Alan Turing Institute

Doh-Shin JEON
University of Toulouse Capitole, Toulouse, France

Gerard LLOBET
CEMFI and CEPR

José Luis MORAGA
Vrije Universiteit Amsterdam and University of Groningen

Contributors

Martin PEITZ
University of Mannheim, CEPR, CESifo and ZEW

Antonio PENTA
ICREA - Research, Universitat Pompeu Fabra
and Barcelona GSE

Michelangelo ROSSI
Universidad Carlos III de Madrid

Pablo RUIZ-VERDÚ
Universidad Carlos III de Madrid

Juan Manuel SÁNCHEZ-CARTAS
Universidad Politécnica de Madrid

Mateo SILOS
OFWAT

Joel WALDFOGEL
University of Minnesota

Introduction

The internet and the technologies that it has spawned are behind the current digital revolution that is progressively changing our lives and the whole economy. In 2006, of the ten largest firms in the world in market capitalization eight were in the energy and financial sector. By 2016, five of the largest ten firms (including the biggest three, Apple, Alphabet and Microsoft) were in the information technology sector and only one energy and financial sector company remained in the top ten.[1] This change is a reflection of the impact of the innovations that new platforms, machine learning, or the sharing economy are leading to in most markets. This book provides a vision from the economic perspective of this digital revolution.

The aim of this book is not to describe the phenomena that this revolution has brought about but to focus on the challenges that the disruption due to the digital economy is likely to generate. We identify three kinds of challenges. The rise of the new superstar firms like Amazon, Alphabet (Google), Microsoft, or Facebook has created concerns for worldwide regulators and competition authorities alike. It is widely believed that these new markets are winner-take-all and if these large firms are left untamed they will transit towards new monopolies. On the other hand, these semi-monopolies (*e.g.* Google controls about 90% of the search market in Europe) are very different from the mammoths of the nineteen and twenty century like Standard Oil, IBM, Ford, or General Motors. They own few physical assets for their level of capitalization, they are not protected by the standard entry barriers like scale economies, and some of the newcomers, like Spotify or Uber, have a large consumer base but, still, they incur in large losses. In addition, many of the digital services are free for consumers or more exactly apparently free, since they are used to gather information about their habits and preferences.

This last change is also one of the sources of the second challenge that the digital economy is likely to pose on the society of the future. The interaction between consumers and firms is undergoing a big change. Information is the new gold of the digital age. Their exploitation using machine learning techniques is creating a tension between providing better and cheaper services to consumers and the protection of a privacy that, until now, was taken for granted. The information generated on consumer preferences also allows a better match with new products and services, permitting niche markets to flourish. The long-tail phenomenon that it is typically associated to cultural goods is spreading throughout the whole economy. The development of review and recommendation systems is providing increasing levels of transparency

[1] "The Rise of Superstars", *The Economist,* Sep 17, 2016.

and reducing the adverse selection and moral hazard problems that used to be prevalent in some markets. As a result, the sharing economy, which was initially related to the sale of products, is now spreading into personal services like short-term housing or car sharing.

New technologies are also changing the way that economic agents are interacting with each others. The usage of digital currencies, like Ethereum or Bitcoin, and the investment in crowdfunding platforms are an important threat for the current role that financial institutions are playing in the economy.

Finally, the way that economics as a science has dealt with most of these issues in the past is in the process of being reconsidered. The standard paradigm is moving from a situation in which the interaction among agents is limited by geographic boundaries, scarcity of information, and high transaction costs towards a new reality of global markets, endogenous and huge amounts of information and decreasing transaction costs. This transition will involve methodological challenges on how to model these new phenomena and how to process information and also a change in focus away from the standard models of competition.

In the present book we aim to shed some light over the previous issues by bringing the contribution of some of the leading scholars in the new fields spawned by the digital economy. We have organized their works in four parts that we develop next.

We start by discussing the increasing prominence of platforms as the basic building block of new digital business models. Traditionally, products where sold by merchants or intermediaries that took possession of the goods produced by other firms and sold them to consumers (e.g. Department stores, Amazon, etc). The internet has fostered the proliferation of platforms where sellers offer their goods to consumers (*e.g.,* Amazon Marketplace, Aliexpress, etc.). The chapter by **Juan Manuel Sánchez-Cartas** called "Digital Platforms and Compatibility. An Old Story in the New World" constitutes an introduction to the economics of platforms. It defines and discusses the concept of a platform as a firm that mediates transactions between its affiliated users. These users are subject to network effects and to the firm's market power. The author also discusses one of basic questions, who pays, and how the elasticity of the demand of each side of the market plays a role in the optimal pricing scheme. Finally, this chapter provides a platform taxonomy that can be described by their two limiting cases. On one of the extremes, there are platforms which operate close to consumers and that they nurture on the data they provide. On the opposite extreme there are platforms with no interaction with final users that build the underlying infrastructure in which other platforms thrive. That is, their customers are other platforms.

In the old merchant model, an intermediary certified the quality of the products that it stocked. In the digital economy consumers often buy directly from sellers through the platform. This new business model requires the creation of new ways to find out the quality of the products and services provided. The next two chapters tackle different aspects of this change carried out using reputation and rating systems. The chapter by **Michelangelo Rossi** entitled "Asymmetric Information and Review Systems: The Challenge of Digital Platforms" studies how online contracting is subject to variations of the classical problems of adverse selection and moral hazard. The chapter applies the analysis of informational economics into the digital economy framework and shows how review systems are used in practice to mitigate such problems.

The chapter by **Paul Belleflamme** and **Martin Peitz**, "Inside The Engine Room of Digital Platforms: Reviews, Ratings, and Recommendations" elaborates on the previous topics. They discuss the impact of review, recommendation, and rating systems. One of their most interesting aspects is that these systems generate network effects. The more people use the platform the more useful the reviews are and more reviews are provided by users. These network effects are often platform specific. As a result, they create a winner-takes-most effect. They study the incentives for platforms to provide informative systems and discuss whether their interests are aligned with social welfare maximization. Finally they explain how recommendation systems help niche firms by generating more visibility for the long tail.

The second part of the book is devoted to pricing mechanisms and search. Platforms have effects on pricing beyond the fees that agents have to pay in order to be affiliated to them. They generate information that is used in order to discriminate prices among final users. In our chapter, entitled "Personalized prices in the Digital Economy" we study how the information gathered by these platforms affects the pricing behavior of firms and their implications for (consumer) welfare. On the negative side, more information allows firms to discriminate prices: to offer individualized prices according to the consumer's willingness to pay, extracting more of their surplus. On the positive side, price discrimination allows the sale of the product to be extended to consumers that otherwise would not be served, and it permits firms to design products that match better their preferences. The common wisdom is that when firms enjoy market power the first effect overcome the second ones and the total balance of price discrimination on welfare is negative. Competition alleviates the rent extraction from consumers and might overturn the results. In this chapter we revisit those results and we show that this intuitive rule is not general and it depends on the setting. We also study the incentives for firms to gather information and consumers to provide it.

Internet users are also exposed to ads when they visit a website. These ads are consumer specific, based on the information gathered by the platform through the usage of cookies (*e.g.,* Google, Facebook). This is a lucrative business and 86% of Google's revenue ($111bn)[2] comes from ads, mostly allocated using auction mechanisms. The chapter by **Francesco Decarolis**, **Maris Goldmanis** and **Antonio Penta,** entitled "Recent Developments in Online Ad Auctions" is an introduction to the economics of digital auctions. Advertisers bid for the placement of their ads either as the result of consumer search (keywords) or as display ads. The paper describes how the auction formats used by platforms have evolved over time. The authors focus on the two more successful auction designs, the Generalized Second-Price Auction (used by Google) and the Vickrey-Clarke-Groves Auction (used by Facebook). They explain the trade-off between the two. The former generates more revenue and the allocation rule is easier to understand but it is strategically complex. The latter provides an efficient allocation and it is strategically simple since it is optimal to bid according to the agent's valuation. The chapter ends with a discussion of the open questions in the design of digital auctions.

The last chapter of this second part of the book is written by **José Luis Moraga González** and entitled "Consumer Search in Digital Markets". The digital economy has not only decreased search costs but it has also affected the way consumers search for products. This change has had an impact on competition among firms that are now concerned about how they can direct consumer search towards their products, for example, through changes in their prices. This chapter analyzes the new patterns of consumer search that have emerged with the digital economy and derives the main implications for competition policy and welfare.

The third part of the book analyzes some of the new digital business models. **Carlos Bellón** and **Pablo Ruiz-Verdú** in their chapter "Crowdfunding: What do we know?" study this new form of financing new projects. Compared to the standard bank financing, crowdfunding provides a new way to deal with the classical maladies of corporate investment. Firms face large uncertainty about the success of their project and information is asymmetric between financiers and entrepreneurs giving raise to adverse selection and moral hazard problems related to the misuse of the funds borrowed. Crowdfunding platforms may alleviate some of these problems. For example, these platforms facilitate the aggregation of disperse information across small investors. More importantly, they make the financing decision contingent on the outcome of this aggregation process, reducing the overall risk and cost for financiers. This chapter reviews the main contributions in this new and growing literature.

[2] *The Economist* "Give me a break" (February 17, 2018).

Online content platforms are discussed by **Luis Aguiar** and **Joel Waldfogel** in their chapter "Digitalization and the Content Industries". The internet has changed the nature and borders of many markets. However, few of them have been capsized in the way that the production of music and media contents have been. These industries have moved from a business model based on content sale to subscription models for consumers and new types of contracts with artists that encompass concert and merchandising revenues. This change has generated in the short run a decrease in revenues for content producers but, at the same time, by facilitating the access of consumers to new content, it has increased the diversity of the supply. This chapter provides an assessment of the global effect of this revolution using recent empirical studies.

Traditional mobility market models are also under threat by the digital economy. Platforms like Uber or Lyft have become powerful competitors to the well entrenched taxi industry in many cities. The chapter "The Economics of the Gig Economy-with an Application to the Spanish Taxi Industry" by **Mateo Silos Ribas** studies this change and explains the technological improvements that these newcomers have introduced. He explains how new technologies overcome the classical market distortions that have been used to justify the protection that the taxi sector has enjoyed in the past. He also uses the case of Spain to provide a sense of the magnitude of the consequences of maintaining the current taxi regulation. He estimates the cost of these regulations in Spain to be as high as 324 million euros a year.

The digital industry has also had a broader impact on society beyond economics. It has modified the way in which news are generated and consumed by readers. The chapter by **Doh-Shin Jeon,** entitled "The Economics of News Aggregators," analyzes one of the most relevant aspects, the emergence of news aggregators like Google News. These aggregators provide consumers with a sample of the news from several sources and are tailored to their interests. The economic literature has identified two opposing effects that these new intermediaries may have on the market for news. On the one hand, they generate a business-stealing effect as some potential readers are satisfied with the information samples provided by an aggregator and do not visit the newspaper. On the other hand, there is a market-expansion effect, as consumers are exposed to competing newspapers that they would not otherwise have visited. The empirical literature indicates that the second effect typically dominates. This chapter reviews the literature both theoretical and empirical and provides policy recommendations.

The last part of this book is devoted to the analysis of new technologies. The first chapter by **Stephen Hansen** is entitled "Machine Learning for Economics and Policy." Machine learning uses algorithms to uncover patterns in data

allowing computers to perform complex tasks. This area has grown in recent years due to the exponential increase in availability of data and increasing processing power of computers. This technology is behind self-driving cars or speech recognition systems. This chapter provides and introduction to this field, explaining supervised and unsupervised learning and discusses some applications to economic measurement and forecasting.

One of the most controversial developments in the digital economy in recent years has been the growing prominence of cryptocurrencies and most specially of bitcoin. Economists do not agree over the potential impact of these new virtual currencies in the financial sector as well as their potential effect over the whole economy. Bitcoin has increased drastically in value but, at the same time, it has been criticized for its volatility, the opacity it allows, and the high power requirements that the mining of new currency requires. The debate about this currency has hidden the main technology that has made the bitcoin and other cryptocurrencies possible: the blockchain. The chapter by **Guillaume Haeringer** and **Hanna Halaburda,** entitled "Bitcoin: A Revolution?" explains how crytocurrencies work. It also provides an introduction to the blockchain technology that it is behind them and it analyzes its potential for other applications like smart contracts.

The final chapter of this book by **Adina Claici** "Big Data and Competition Policy" discusses how the massive use of data by firms is likely to modify market competition and how competition authorities have intervened until now. Because markets in which data usage is massive also tend to be concentrated, the first question is whether data constitutes a barrier to entry or not, preventing competition from arising. This question has implications for merger decisions as shown in the case of Whatsapp and Facebook which is discussed in the chapter. Big data has also implications for the potential of large firms to abuse their dominant position. This chapter provides a thorough discussion of this risk using the Google Shopping case. Finally, it analyzes how the use of data can facilitate collusion among firms by, for example, using unsupervised machine-learning algorithms.

Madrid, July 2018

Juan-José Ganuza and **Gerard Llobet**

PART I

Platforms and Information

DIGITAL PLATFORMS AND COMPATIBILITY. AN OLD STORY IN A NEW WORLD

Juan Manuel SÁNCHEZ-CARTAS

Abstract

Digital platforms communicate with each other. They exchange data about their customers using common telecommunication protocols that create compatibility networks among platforms. However, the use of data is not homogeneous, some platforms freely share their data, and others sell data. In this work, we study the role of data sharing among platforms, and how this behavior affects traditional economic insights. We describe the role of data in the new generation of digital platforms, how the old economic insights still apply in some cases and the new behaviors that are exclusive of digital platforms. Lastly, we contextualize our findings by analyzing the fitness-tracker market.

Key words: Compatibility, digital platforms, fitness-trackers, digital competition.

JEL classification: L10, L15, L86.

I. PLATFORMS ARE CHANGING OUR LIVES

Sunday, 9 am. Susan begins to warm up. She turns on her iPod. She has more than 100 songs, and everything is well organized in her playlists. She knows that a recent hit has just released, and she had synchronized her iPod with Spotify the previous night. But before going out, she takes a look at her wrist. Her Garmin is on and says that the heart rate monitor is ready. She can start running. Lastly, she checks out her phone, she wants to record the path but also, she wants to receive live updates from her friend Eva, who is already running nearby.

Susan is doing what she does every Sunday. This routine is made automatically. It is so normal as it is the warm up. But Susan is not a *technophile,* she is a normal girl. But platforms have become an integral part of her life. And she is not the only one. We live surrounded by platforms. They are everywhere, and they are disrupting businesses, behaviors and even governments. This revolution is based on allowing interchanges, transactions, and connecting people. But thanks to the information and communication technologies (ICTs), the consequences of these interchanges are global, and they are changing how we buy, communicate, and even run.

The idea of putting in touch two or more groups of people who need each other is not new. Newspapers, academic journals and even fairs work in this way. They "connect" readers and advertisers, researchers and readers, and buyer and sellers. However, ICTs have allowed us to scale up this idea to the whole world. Traditional newspapers or fairs have two clear shortcomings that digital platforms avoid: the physical copy and the physical presence. To benefit from a fair, you have to be there. To read a paper, you need a copy. In both cases, it is costly to print a newspaper or to set a stand at a fair. However, digital platforms allow us to overcome these two issues: you do not need to be physically in some place, and copies can be made for free. The same "message" can be delivered to millions at almost no cost.

These two features have allowed platforms to reach global significance. The larger the number of users, the more relevant they become. All platforms are made by their users. Amazon is made by sellers and buyers, Facebook is made by users and advertisers, Youtube is made by watchers and broadcasters (youtubers), etc. When in history has a service reached such relevance worldwide? This is the first time. And it is a revolution. New behaviors, jobs, services, regulations and so on are starting to emerge worldwide.

Youtubers and influencers have become more relevant for promoting products than celebrities.[1] New services have appeared in cities competing with traditional services such as Uber and Lyft. They have become a concern to regulators who observe how platforms are using gaps in the legal system to generate new businesses.[2]

But these platforms are not isolated events. They tend to be related with one another by complex networks. Some of them are built on top of other platforms that are used as benchmarks, like Android or iOS. Others are creating new ecosystems on top of those platforms such as Garmin or Facebook. And others are creating complex networks by which users in some networks can send data to other competing networks. Compatibility allows us to create new platforms on top of previous ones using common communication protocols but also, it allows us to send and receive data from other platforms, partners and competitors alike.

How do platforms interchange data is a major issue for regulators. Global platforms such as Facebook or Google have created vibrant ecosystems full of users and developers that are generating huge amounts of data that they interchange. But, to what extent the use of data in these networks influences our economic intuitions? Can we rely on our traditional insights? Or is this time different?

Let us follow Susan once more to see how platforms are using your data, and how compatibility is changing the competition in these markets.

II. ECONOMIC PLATFORMS. WHAT IS THAT?

Up to now, we have talked about platforms, but we have not defined them. Let us take a moment to study how the academic literature has defined a platform from an economic perspective. What is a platform? There are multiple definitions depending on the point of view (engineering, computer science, economy, etc.). But we are interested in economic platforms, also known as multi-sided platforms. In a nutshell, a multi-sided platform is a service that "coordinate[s] the demand of distinct groups of customers who need each other in some way" (Evans, 2003).

[1] https://www.thinkwithgoogle.com/consumer-insights/youtube-stars-influence/

[2] https://www.mwe.com/en/thought-leadership/publications/2016/03/according-to-paris-court-of-appeal-jurisdiction

Unfortunately, multi-sided platforms[3] do not have a clear and widely accepted definition as it has been pointed out by van Damme *et al.* (2010), Evans (2011) or OECD (2009). In fact, *you know a [multi-sided] market when you see it,* see Rochet and Tirole (2006).

Its identification presents several problems. On the one hand, we have to define what we mean by "platforms" because there is no "industry of platforms" in official statistics. In fact, platforms are technologies that can be used by a great number of industries, (see Evans, Hagiu and Schmalensee, 2008). In this case, we can consider that a platform is a technology (or a procedure) that minimizes transaction costs, or a technology that creates a value allowing transactions that otherwise would not occur, (see Evans and Schmalensee, 2005). Nevertheless, this definition is very broad and, virtually, every market could be studied as a particular case of multi-sided markets. The term "two-sided markets (platforms)" was first used in Rochet and Tirole (2003). Nevertheless, these models had been studied before by Parker and Van Alstyne (2000), Caillaud and Jullien (2001) and Caillaud and Jullien (2003). In these last two cases, they refer to the platforms as intermediaries (or "cibermediaries" in their own words).[4] Initially, Rochet and Tirole proposed a definition that considered markets and platforms as the same item. Their definition stated that a platform was two-sided if the number of transactions on the platform can be influenced by changing who pays more and who pays less. In that case, we face a platform.[5]

For example, in the credit card market, buyers normally do not pay for the transaction, but sellers do. If we evenly share the price paid by sellers among sellers and buyers, the number of transactions will not remain equal. Fewer buyers will be willing to pay with credit card, and fewer sellers will accept credit card too. The main shortcoming of this definition is that it only relies on the price structure and on considering markets in which platforms can control the transactions like credit card markets. However, they do not take into account

[3] For some authors "multi-sided markets" and "multi-sided platforms" are not the same because there are important normative implications. For instance, Evans and Schmalensee (2013) are against the use of the terms "two-sided markets or "multi-sided markets" because they think that *multisideness* is an attribute of individual companies. It does not need to be an attribute of every company in the market. For example, in the rental car industry, there are intermediaries that put in contact renters and drivers, they behave like two-sided platforms, but in the same market, there are renters who get in contact with drivers directly, and they are not two-sided platforms.

[4] The birth of this literature is a conflictive issue because, for some authors, the birth is when the term "two-sided market" is coined. To others, it is when the first paper with inter-dependent demands between two sides was published. In this regard, the birth is attributed to Parker and Van Alstyne.

[5] *A market with network externalities is a two-sided market if platforms can effectively crosssubsidize between different categories of end users that are parties to a transaction. That is, the volume of transactions on and the profit of a platform depend not only on the total price charged to the parties to the transaction, but also on its decomposition, Rochet and Tirole (2003).*

markets like newspapers, where the platform (newspaper) cannot control if the reader is interested in the advertising.

One of the first works in proposing a broader definition was Evans (2003): *Multi-sided platforms coordinate the demand of distinct groups of customers who need each other in some way.* In contrast with the Rochet and Tirole's definition, Evans' considers the possibility of platforms that do not control transactions. The main shortcoming of this definition is that it is too broad. Almost every relationship may fit the Evans' condition of "who need each other in some way".

On the other hand, the great contribution of Rochet and Tirole is to highlight the difference between one-sided and two-sided markets. In other words, what really matters is who pay for the service. Their definition emphasizes the essential role of indirect network effects. For example, let us consider a nightclub in which men's ticket is 10 euros and women's ticket is 5 euros. The total price paid by both sides is 15 euros but, if we evenly share the price (7.5 euros each), will there be the same number of customer in the nightclub? If the answer is no, that is a hint that we are facing a two-sided platform.

Rochet and Tirole recognize that under their definition almost every company would be a two-sided platform. However, they argue that, at least in competitive environments, companies are often *de facto* one-sided platforms because if the number of companies tends to infinity, the networks effects tend to zero, *i.e.,* without network effects, there is no multi-sided platform. The larger the number of platforms, the less likely we will deal with a two-sided platform.[6] However, the vast majority of the literature uses a simpler and straightforward definition (also highlighted by Rochet and Tirole), the presence of indirect network effects: *the net utility on side "i" increases with the number of members on side "j".* In general, a lot of definitions are based on the existence of these externalities, such as those in Evans (2003), Schiff (2003), Wright (2004), Ambrus and Argenziano (2004), Hagiu (2004), Jullien (2005), Anderson and Coate (2005), Parker and Van Alstyne (2005), Armstrong and Wright (2007), Parker and Van Alstyne (2005), Evans, Hagiu and Schmalensee (2008), Weyl (2010), Weisman and Kulick (2010), Ivaldi, Sokullu and Toru (2011), but this idea is not shared by all authors.

[6] From an economic point of view, the interesting feature is the link between their definition and the Coase Theorem. The Coase Theorem states that if property rights are clearly established and tradeable, and if there are no transaction costs nor asymmetric information, the outcome of the negotiation between two or more parties will be Pareto efficient, even in the presence of externalities. The Coase's idea is that if outcomes are inefficient and nothing hinders bargaining, people will negotiate their way to efficiency. In the previous example, couples can reallocate their tickets. A nightclub in which only couples go would be a one-sided platform. In the credit card example, sellers and buyers cannot coordinate themselves to reallocate their prices, so the Coase Theorem fails. Therefore, this market is more likely to be a two-sided one.

Hagiu and Wright (2015) criticize Rochet and Tirole's approach and they proposed a definition of multi-sided businesses based on two characteristics:

- Multi-sided businesses enable direct interactions between two or more sides.
- Each side is affiliated with the platform.

By "direct interaction", they mean that two or more sides retain control over the essential terms of the interaction. For example, on the Uber platform there are two sides, users and drivers. Drivers retain control rights over the car (it is the drivers' car) as opposed to the one-sided intermediaries (taxi companies) that have total control over their fleet. Therefore, this is the main difference between the one-sided and the multi-sided worlds. By "affiliation", they mean that users on each side consciously make platform-specific investments that are necessary in order for them to be able to interact with each other directly, for example, paying membership fees or registering. In the Uber example, both users and drivers have to invest time in registering in the App. The affiliation helps to distinguish multi-sided platforms from inputs suppliers.

The most remarkable contribution by Hagiu and Wright is that their definition does not require any reference to indirect or cross-network effects. Hagiu and Wright consider they are neither necessary nor sufficient to define a multi-sided platform. However, indirect network effects could be consequence of "affiliation" or "direct interaction". The authors consider that Rochet and Tirole's hypothesis about every market with indirect network effects being a two-sided market is not correct, and they explain it in this way: *note that indirect network effects are not limited to multi-sided platforms [...]. [In] traditional consulting firms, clients will be attracted to a consulting firm that has many other clients since this means it will have access to a greater number of qualified consultants.*

Given the complexity of defining a two-sided market, it is normal to find works that consider different definitions. Some authors such as Filistrucchi and Klein (2013) or Evans, Hagiu and Schmalensee (2008) have shown that reality is very ambiguous. In fact, Filistrucchi and Klein (2013) and Rysman (2009) claim that, theoretically, Rochet and Tirole's definition can include one-sided cases. Another point of criticism related to the Rochet and Tirole's and Evans' definitions is that all of them refers to "markets", not to businesses or platforms like the Hagiu and Wright's. Rysman and Evans share this criticism. They point out that the definition of multi-sided markets is not totally correct because it is hard to find "pure multi-sided markets". On the other hand, it is easier to find "multi-sided businesses/platforms". We can find markets where there

are companies using multi-sided strategies and companies using one-sided strategies.

Rysman (2009) uses as an example Amazon that was one-sided in the market of books and multi-sided in other markets. That is why it is important to Rysman to focus on the strategies adopted by firms because *multisidedness* is an endogenous decision of firms. The main question is not to know if a market is a multi-sided one, virtually all markets might be multi-sided to some extent. What is relevant is to know how important multi-sided issues are.

Highlight 1. There are many definitions of multi-sided platforms. And many of them use the terms "platform" and "market" interchangeably. Nonetheless, almost all of them emphasize the role of a technology enabler (the platform) to mediate between the transactions of two or more sides.

In general, the vast majority of authors and international organisms recognize that there is not a universally accepted definition of multi-sided markets or platforms yet. There is a consensus on the idea of two or more groups of agents who need each other in some way and who rely on platforms to intermediate transactions between them. There is also consensus on the idea that it is more important to determine the linkages between the two sides of the market than the market itself, (OECD, 2009; Filistrucchi, Geradin and Van Damme, 2012; or Weyl, 2010). Weyl highlights that definitions have their flaws but, in general, multi-sided markets have three features:

- There is a multi-product firm. A platform provides distinct services to two sides (or more) of the market.
- There are cross network effects. Users' benefits from participation depend on the extent of user participation on the other side of the market.
- Bilateral market power. Platforms are price setters on both sides of the market.

The author argues that the failure of any of these conditions makes simpler and better understood other models. If a platform does not explicitly charge different prices to different groups of users, it is best viewed as a standard, one-sided company. Obviously, the role of a platform will depend on the market where it is operating. In summary, definitions of multi-sidedness are controversial. There is no consensus. However, as it is pointed out by Filistrucchi, Geradin and Van Damme (2012): "Although, at first sight, it appears to be

still some debate on the exact definition of a two-sided market, the different definitions proposed appear to be consistent enough to allow the practical identification of two-sided markets."[7]

Highlight 2. There is no consensus about the definition of multi-sided markets. Nonetheless, the practical identification is consistent with the idea of two or more groups of agents who need each other in some way and who rely on platforms to intermediate transactions between them.

1. Pricing Platforms. Who Pays?

What makes interesting and different multi-sided platforms is the way they set prices. Previously, we have seen the example of the nightclub that sets a different price for men and women. This asymmetric pricing scheme is the main characteristic of multi-sided platforms.

Platforms realize that some groups of consumers value more the presence of other different groups of consumers (indirect network effects). For example, readers and advertisers, men and women, buyers and sellers, etc. However, some consumers value more the presence of others types of consumers than the other way around (for example, on average, men may value more the presence of women than women the presence of men). In this situation, platforms find profitable to reduce the price on one side (women) to increase the number of those consumers, and to attract more consumers on the other side (men). In summary, multi-sided platforms tend to set an asymmetric price structure in which one side is the profitable one, and the other one is the loss side.[8] This asymmetric price schema is common in markets like credit card markets. Sellers have to pay a fee per transaction while users do not pay such fee. Another example is media platforms. Free newspapers or free TV programs are free because, in that way, they are able to charge higher prices to advertisers.[9]

This asymmetry in prices is due to the indirect network effects. And it creates a great challenge because it breaks some traditional rules about pricing.

---

[7] See Sánchez-Cartas and León (2018) for an extensive review on multi-sided markets.

[8] I am aware that pricing multi-sided platforms is far more complex than the description I provide here. Nonetheless, explaining the different pricing policies that may arise in these markets is far beyond the scope of this work. See Rochet and Tirole (2004), Rochet and Tirole (2006), Weyl (2010) and Cabral (2011).

[9] Rochet and Tirole named this behavior "the seesaw principle," and they define it as follows. *A factor that is conducive to a high price on one side, to the extent that it raises the platform's margin on that side, tends also to call for a low price on the other side as attracting members on that other side becomes more profitable.* Later, Weyl (2010) stated that the seesaw principle was the most robust result on comparative statics of two-sided markets.

Highlight 3. Multi-sided platforms tend to set an asymmetric price structure in which one side is the profitable one, and the other one is the loss side.

For example, Evans (2003) points out there is a disconnection between prices and marginal costs. This feature contrasts with one-sided markets in which there is a clear relationship between the prices and costs. Evans (2011) argues that it is possible that a platform will respond to an increase in costs on one side with an increase in prices on the other side. Regarding this relationship between prices and costs, Jullien (2005) argues that, in multi-sided platforms, it is common to observe prices that are unrelated to marginal costs. From a social point of view, Rysman (2009) points out that: *Theoretically, it is often hard to establish whether a given price in a two-sided market is higher or lower than socially optimal, or even whether greater competition would make the existing price rise or fall.* This contrasts with traditional markets in which it is traditionally believed that more companies imply more competition and more welfare.

Highlight 4. Prices in multi-sided platforms tend to be disconnected from costs. Even the prices that are socially optimal can be unrelated to costs. This is a consequence of the indirect network effects. Optimality calls for subsidies from one side to others. Neither prices above costs are always a signal of market power nor prices below marginal costs are a signal of predation.

In multi-sided markets, we can find two types of prices: membership fees and transaction fees. The first ones make reference to the price that a user pay for entering the market. For example, the price paid by readers to access a digital newspaper. The latter ones make reference to the price paid each time that a transaction occurs. For example, the commission paid by a vendor when a buyer pays by credit card. Both fees can be found together in some markets. For example, a digital streaming platform may have a monthly subscription, but to access specific content, you have to pay an additional fee for each minute you use that content. The choice of fees is not easy, and it depends on the control that the platform has over the transactions, the information about the users, the market, incumbents, consumers' perceptions, etc.[10]

Nonetheless, one interesting feature is the static nature of prices in multi-sided platforms. Prices do not change, at least with regard to their structure. Once the

[10] The correct choice of fees is beyond the scope of this paper. See Filistrucchi (2008), Rochet and Tirole (2006) or Weyl (2010).

platform is stable, prices tend to be stable, see Evans, Hagiu and Schmalensee (2008). However, the nature and structure of those prices can have different origins depending on how the value is created in the platform.

Lastly, although multi-sided platform prices seem to be quite different than traditional prices. They have common aspects. For instance, the higher the differentiation among platforms, the higher the prices on at least one side. Hagiu (2004) and Evans (2002) find that differentiation guarantees the existence of several platforms in the same market. Rysman (2009) summarizes this feature as: "if [platforms] can differentiate from each other, they may be able to successfully coexist."

Highlight 5. Although multi-sided platforms set prices that are quite different in their structure from those in a one-sided market. There are some ideas that remain valid. For example, the higher the differentiation, the higher the prices on at least one side.

Once that we know what a multi-sided platform is, and how different are their prices, let us return to Susan and her daily activities.

III. THE PLATFORM REVOLUTION: A CLASSIFICATION BASED ON THEIR RELATIONSHIPS

How different is running nowadays! Just after finishing running, Spotify knows which songs Susan listened to Spotify also knows that she was running because she has her Facebook account linked with Spotify, and she has already posted her route. Also, Garmin has just confirmed her GPS position during the route, her heart rate, her speed, and the comparison with her friend Eva, but Garmin is not the only one. Google Maps also knows that she was running in the park near her home, and MyFitnessPal also knows her heart rate, weight, and speed because Susan likes to control how much calories she burns and she has linked Garmin and MyFitnessPal. It was only 30 minutes of workout, but up to five different digital platforms have been involved. All of them related to the same task: running. And all of them related in different ways. Spotify and Facebook share a compatible communication protocol. Garmin and MyFitnessPal another one. And all of them are built upon Android or iOS. In other words, there is a complex network of compatibilities among platforms.

Nowadays, compatibility has different names and implications. It can refer to the compatibility in communications protocols among platforms. This

is the classical compatibility definition by which several devices, products or items can be used together as a single device, product or item. However, in digital platforms, there is a new way of compatibility that refers to the use of data. It broadly refers to the access to the competitors' networks, or to the use of competitors' data. This case has other names such as shared networks, shared databases, synchronization agreements, data sharing agreements, etc. Although each one of those names may have different practical implications (different degrees of access to the databases, protocols, etc.), it is obvious that all of them refer to the possibility of accessing competitors' data. In this sense, it is important to address the relationships among platforms properly. To do so, we need a way to classify and differentiate platforms and their relationships.

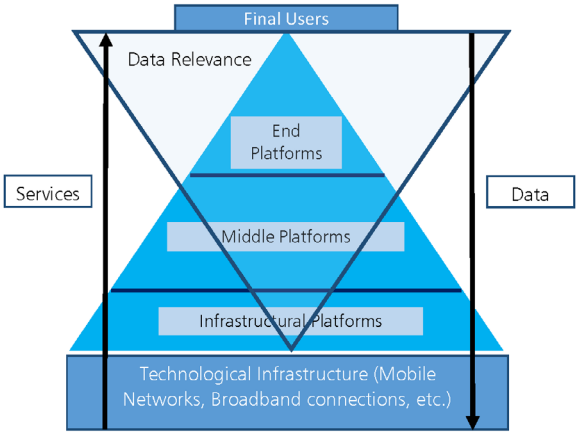
Some platforms are more subtle than others, and it is not easy to notice them, but they are everywhere, and you cannot escape from them. In any daily activity, there are at least three platforms involved. The first ones are infrastructural platforms. They are essential in any modern device, such as smartphones. They are the operating systems (Android, iOS, etc.), and they offer you the basic functionalities that make your phone "smart", but they record information about your phone activity, use of the internet, etc. that can be used by third party companies to develop new services. The second ones are the middle-platforms[11] that are platforms built upon the previous ones, but they also have other services or platforms built upon them. For example, Facebook or Google Maps are two middle-platforms because both are built upon an operating system, and both of them are used as a benchmark for other services or platforms such as Facebook games or mapping services. These middle-platforms offer a service to users but they also offer the possibility of building new platforms upon them. Lastly, we have the "end-platforms", that are apps built upon all the rest of layers. They can be platforms as well, but nothing prevents them from becoming simple apps, in the sense that they only offer a service to users and do not worry about creating an ecosystem of other apps around them. Examples of these platforms are Whatsapp, Imgur or Shazam.[12]

These three layers are related, and each one is built upon the others. Clearly, there is a vertical chain that links those platforms. The only way to use a Facebook game on our smartphone is to run Facebook on Android, iOS, or other operating systems. These operating systems provide a basic environment for other platforms. In Figure 1, we can observe a scheme of these relationships. If we consider Susan's workout again, we can relate each platform to a category in our previous taxonomy. Spotify is an end-platform. Users use it for listening to music, which is the main service of the platform. However, Garmin

[11] Do not confuse them with middleware, or middleware platforms.

[12] It is true that some of those platforms can become middle-platforms. The differentiation among them is subtle, and it mainly depends on the use of each user.

FIGURE 1
VERTICAL RELATIONSHIPS



or Facebook are middle-platforms. Both platforms have an ecosystem around them with other platforms or apps that are built on top of them. Nonetheless, this classification depends on the specific use of each person. If you only use the Garmin or Facebook main platforms and none of their third-apps, then you use them as end-platforms. Lastly, Susan was using her smartphone and her iPod, which run on Android and iOS respectively, infrastructural platforms.

Highlight 6. There is a complex set of relationship between all the digital services we daily use. Many of those services are platforms, and they may play three different roles. The infrastructural one (the basic functionalities), the middle one (enablers), or end one that only offer a service to end-users. Each one is built upon the previous one.

1. Platform Relationships and Data: A Chain of Platforms or Nested Platforms?

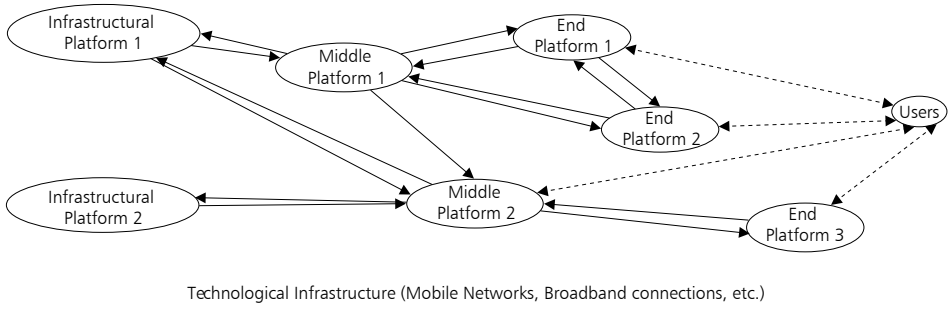
As we observe in Figure 1, services are built upon platforms, and the relevance of users' data decreases when we move towards the bottom. This representation allows us to depict not only the relationships among platforms but also, the relationship with users' data. However, this is not the only possible representation. There is another different one. In Figure 2, we observe another potential representation of the relationships among platforms. The

representation of platforms and their relationships is quite relevant because each representation emphasizes different features of those relationships.

For example, Figure 1 creates the idea of vertical relationships, a value chain. However, the Figure 2 suggests a nested vision in which all the elements belong to a technological environment where platforms and customers are linked by bilateral relationships (services-data/prices). These different visions of the same problem spark different interpretations of the relevance of compatibility among platforms (and the use of data).[13] For example, Figure 1 may suggest that there is a value chain, and maybe the users are paying expensive services because everyone in that chain is trying to earn profits (in economics, this is known as the double-marginalization problem). On the other hand, Figure 2 does not suggest that issue, but it suggests that some platforms are more relevant than others because of their relationships with other platforms in their market, or in other markets. Also, Figure 1 suggests a linear transfer of data, from the top to the bottom. On the other hand, Figure 2 suggests that transfers can go in any direction.

FIGURE 2

NETWORK RELATIONSHIPS AND INTERCHANGES OF DATA



But the relationship among the platforms is not the only relevant topic. It is also relevant the relationship of platforms with their customers. We need to consider that digital platforms provide services to several sides, and all the sides have to be taken into account. If we consider one side only, it may lead to wrong conclusions. For example, Susan is worried about the use of her data by MyFitnessPal. If we want to study the use of her data, we cannot focus on the relationship between Susan and MyFitnessPal only. We have to consider the

[13] A similar taxonomy with horizontal and vertical relationships among platforms can be found in Belleflamme and Toulemonde (2004).

relationship between MyFitnessPal and developers, or other companies that use her data.

Lastly, it is necessary to consider that when we refer to compatibility as shared networks or shared data, there are different uses of that data that lead to different conclusions. In Figure 1, we observe the first relationship. The vertical relationship, in which platforms provide other platforms with data. In this case, data is a mere input to produce an output. This is the most intuitive case. Platforms generate a lot of data that they sell to third-party companies that benefit from it. It is the digital equivalent to buy a hammer in an appliance store and to use it to build a closet or a rack. You use it as an input to produce an output (the closet or the rack). In Figure 2, we observe the second relationship. The horizontal one. It suggests another type of use of data in which platforms share data not only with third-party companies but also, with competitors. In our previous example, this relationship implies that you ask your competitor to share with you the hammer, and both build closets or racks using the same hammer. This situation is strange for most of us, but it is common in digital platforms. In the following sections, we analyze these different uses of data and those different relationships among platforms.

Highlight 7. There are different ways of representing the relationships among platforms. Depending on which topic we want to address, some of them are more useful than others. In the case of data, each representation points out different roles of data. The vertical relationship highlights the role of data as input in the value chain, and the horizontal relationship highlights its role as a link with competitors and other stakeholders.

2. Other Classifications

The previous classifications are not the only ones that we can find in the literature. There are a lot more. However, the previous one allows us to focus on the relationships that platforms have with other platforms. Other types of classifications do not allow us to address such relationships. For example, one of the most interesting classifications is the one proposed by Filistrucchi (2008). He classifies two-sided markets in two categories

- *Media type,* these platforms are characterized by the absence of observable transactions. For example, TV channels or newspapers. In these cases, advertisers display ads, but they do not know if someone is influenced by those ads. These markets are also characterized by setting "membership fees" only.

- *Payment card type,* these platforms are characterized by the observability of a transaction between the sides, like a payment with a credit/debit card. The platform can monitor the transaction, and it can apply transaction fees.

This classification is quite useful. It only requires knowing the pricing policies. However, it does not allow us to classify platforms with respect to their relationships.[14]

There are also other classifications that can be interesting such as the ones proposed by Evans (2003), or Tiwana (2013). However, they are not useful for illustrating how data influence the behavior of platforms. That is why we omit an extensive analysis of those taxonomies. Nonetheless, there are other interesting classifications that link platforms to their pricing strategies. For example, Rysman (2009) points out that it is normal to find multi-sided platforms and traditional re-sellers. We also observe in digital markets how platforms compete with traditional business models. For example, Uber and taxi companies, or Amazon pantry and supermarkets.[15]

IV. COMPATIBILITY AND THE USE OF DATA IN DIFFERENT MARKETS

Susan has a smartphone and has total control of her life with it. Almost any uncomfortable task that she used to do ten years ago is easy to do using her smartphone nowadays. This morning, she was in a hurry. She was rushing to the bus stop while she was checking out the weather. Today, it will be sunny. She also used another app that estimated that the bus would arrive in 5 minutes. She was on time. However, when she was on the bus checking out Facebook, she realized that she forgot her food at home. She opened the advertisement that she saw on Facebook about HealthyOut, and she placed an order to deliver Chinese food at her work at 12 am. Because she is worried about her nutrition, she shared the calories information with her MyFitnessPal account. This account is linked to Garmin connect, which quickly updates the information about calories burnt in her MyFitnessPal account. Lastly, because

[14] Nonetheless, expert readers will notice that the literature has focused much more on "media type" markets than on "payment card type" ones. Therefore, the following sections are highly influenced by "media type" markets.

[15] Which business model is better is a topic that is beyond the scope of this work. Nonetheless, there is no a better business model. It depends on the market. See Hagiu and Halaburda (2014) or Rysman (2009) for a discussion on this topic, or see Sánchez-Cartas and León (2018) for an extensive review on other classifications.

she placed the order while using Facebook, this platform and Google (which controls the Gmail account that is related to the phone) also know that she placed that order.

Behind this chaotic set of relationships among platforms, there are three behaviors that involve Susan's data. When there is no data sharing; when data is extracted and used as an input by the platforms; and when platforms share data with competitors. For example, the weather app or the bus app only display advertisements and the information Susan is looking for. There is no further interaction between Susan and the app. This can be considered the simplest case. Also, most of the people think that this is the common pattern. However, we are sorry to disappoint you, but it is not.

The second type of behavior is the most common one. This is the case of the big players in the industry such as Google, Apple or Facebook. All of them have platforms that can extract a lot of information about you constantly. Maybe you do not realize it but, if you have Facebook on your Android (iOS) smartphone, both Facebook and Google (Apple) know where you are, and probably, what you are doing. This case is scary for a lot of people, and it sparks a debate about privacy and customized services. However, we will not address this concern here. This case also encompasses situations where different platforms from different market segments cooperate and make their services compatible. For example, the integration of Youtube on Facebook or the possibility of sharing your Shazam songs on Twitter.

The third type of behavior is the case of data sharing among platforms that compete with each other. This is the most intriguing case because it refers to platforms that compete with each other for the same users and developers, but they "share their data". They offer their databases to competitors. That is the case between Garmin and MyFitnessPal.

Data are the essential good in those relationships. And it is not clear if the traditional economic intuitions remain valid when we consider digital platforms. Wright (2004) points out that conventional knowledge from the classical economics literature may lead to mistakes when addressing digital (multi-sided) platforms. In that sense, some conclusions may not be robust in those markets. This suggests that policymakers have to be careful not to base their policies on inadequate generalizations about markets, especially in ICT markets.

The main difference between traditional economics and platforms is subtle, but it motivates a whole line of research. In the traditional economics, consumers value the presence of other consumers in the services. One example is

the social networks. Users value whether or not their friends are on the platforms. However, in digital platforms, some consumers may value the presence of other consumers too. The essential difference is that they also value the presence of other types of consumers. The social networks are one example of this. Companies value the presence of users, but companies are also consumers of the social network. However, they have a different purpose than people who connect with their friends. To what extent this subtle difference is affecting the behavior of consumers and platforms is an ongoing research topic. Nonetheless, some advances have been made, and we can point out some consequences of realizing that different types of consumers interact with each other in digital platforms. In the next section, we will focus on compatibility and on the use of data generated by that compatibility.

Highlight 8. Data may play three roles: No use at all because data is not "harvested", as an input in the value chain to increase the value of the companies' products or services, or as a commodity that it can be sold or shared with third-parties.

1. Compatibility: Old Rules in New Behaviors

The idea of compatibility that we use today when we refer to digital platforms is similar to the one proposed by Katz and Shapiro (1985). They defined compatibility as follows. *If two firms' systems are interlinked, or compatible, then the aggregate number of subscribers to the two systems constitutes the appropriate network. If the systems are incompatible, such as Telex and cable, then the size of an individual system is the proper network measure for users of that system.*[16]

This definition does not emphasize the role of data, but the role of users who use the same service. There are two situations in which this idea of compatibility can be considered to address digital platforms. First, in the launching phase, many platforms behave like traditional companies, serving only one type of customers, (see Rysman, 2009). The reason is that these platforms do not have enough users to attract other types of customers like advertisers. In this sense, some platforms are born as traditional companies that consider one type of customer only. For example, this was the case in social networks. In the beginning, their purpose was to put in touch friends, family, colleagues, etc. They were focused on attracting users who interact with each other. Then, platforms realized that advertisers value the information about people's relationships but also, they value even more the information about people's

[16] They also defined the hardware-software compatibility such as: *If two brands of hardware can use the same software, then the hardware brands are said to be compatible.*

tastes. Then, the digital platforms as a multi-sided business started. So, during those initial phases, this definition of compatibility is valid because platforms are focused on the number of users only.

The second case is when platforms allow you to use the services of a third-party company to communicate with external agents. For example, Susan shares with her friend Eva all their running routes, her performance, heart rate, speed, etc. Platforms may allow her to send that information to her friend Eva. This option can be considered as a complement to the platform's services. In this sense, this case can be analyzed as a complementary good, and traditional intuitions may apply because we put emphasis on connecting people, not the data[17]. However, if platforms use that data for commercial purposes, these intuitions may not apply, and we have to address new approaches.

If we consider the initial description of compatibility and we omit the comercial use of data for a moment, we can observe how platforms use the compatibility to create value for users. If we pay attention to Figure 1 again, it is crucial that platforms at all the levels complement each other to create value for users. Facebook will have no value at all in a smartphone in which it crashes every five minutes. However, if it runs fast and it is a reliable app because it is built on the top of a compatible system, the bundle: smartphone plus apps is quite valuable. This complementarity among services allows platforms to increase the adoption, and these intuitions are valid for both, digital platforms and traditional businesses.

When platforms allow you to connect with other users on different platforms, the complementarity and the compatibility help to foster the adoption of all the platforms. All those platforms become more attractive because their users' bases become larger. We observe this behavior in digital platform markets such as fitness trackers. In this case, we can consider that companies have allowed compatibility between their devices and the digital platforms of other companies in an attempt to foster the adoption of their products. For example, Eva and Susan use different devices. They would not be able to compete nor to compare their performances if companies were incompatible. In this sense, many companies allow cross-synchronization of their devices with other platforms because a critical mass of users can be reached easily.

[17] With traditional intuitions, we refer to the intuitions derived from the network economics literature. Many works have been developed in the network economics literature, and we do not have time to review all of them. For a comprehensive review of the literature see Economides (1996). For an introduction to the topic, see Belleflamme and Peitz (2015).

In the case of end-platforms, many of these platforms are compatible by default with the infrastructural and middle-platform layers. This is the case of Shazam and Whatsapp, Twitter and Youtube, or the integration of different apps in different devices such as smartphones or wearables. In fact, this kind of compatibility is normal because, if compatibility only requires one side to allow it, it is normal that the one interested in the compatibility will do it (see Matutes and Regibeau, 1988). This compatibility between independent products can boost the demand or the adoption of those products but also, it makes more valuable the product for some users,[18] and more profitable for companies (see Matutes and Regibeau, 1988). However, this compatibility among platforms creates incentives to increase prices because:

1. Compatibility increases the value of the goods (see Farrell and Saloner, 1985 or Economides and Salop, 1992).

2. Compatibility reduces the incentives to compete in prices since the effect of reducing prices affects all the complementary products (see Matutes and Regibeau, 1988).

One example of all those intuitions is the iPod. When Apple made its iPod compatible with PCs, sales took off sharply.[19] After that, iPod prices remained almost untouched.[20] Considering the technological race in these devices and the emergence of other competitors, it seems that the traditional intuitions give us an interesting answer to why prices were high during so much time. Nonetheless, we do not have to forget that other things are happening at the same time that increase and reduce prices such as technology evolution (the increase of prices for new generations) or changes in the tastes of users (the reduction of prices for users who value more new generations than old ones).

But these are not the only intuitions that remain valid in digital platform markets. On the other hand, if we pay attention to the development of the operating systems such as Android Things, Android Wear or iOS. They are formed by different layers that use different standards and protocols that are especially addressed to developers. In those cases, traditional intuitions still apply, and there are many examples of behaviors that can be explained by the traditional economic literature. For example, it is quite common to hold technical conferences for developers from time to time. This literature highlights that, in this way, communications allow to set standards that help in fostering

[18] These features were early highlighted in the literature. See Katz and Shapiro (1985) or Farrell and Saloner (1985).

[19] http://www.ilounge.com/index.php/articles/comments/instant-expert-a-brief-history-of-ipod/

[20] https://www.macworld.com/article/1053499/home-tech/ipodtimeline.html

compatibility among services.[21] The literature also highlights that it is normal to develop systems that are incompatible by definition such as iOS and Android, that represent two different approaches to the same issue: an operating system for mobile devices. In these cases, they are born as incompatible services, but they adopt partial standardization during the evolution of the systems because it is profitable.[22]

However, there are other situations in which these intuitions do not apply. For example, when we deal with the commercial use of data. When Eva and Susan share their performance, they are also sharing data. These data can be sold or can be given for free. This is a consequence of compatibility, and this consequence was not addressed in the traditional economic literature. Recently, we started to pay attention to it. The definition of compatibility is the same than the one proposed by Katz and Shapiro. However, this time is different.

Nonetheless, the impact of the commercial use of data is not homogeneous because it depends on the laws around digital platforms. Countries differ in their laws, and digital platforms have to adapt to them. In this sense, the legislation of each country is essential to fix the business model of each digital platform. For example, Uber works as a multi-sided business in California, but as a traditional taxi service in Madrid, and it is illegal in London.[23] In the European Union, the use of personal data is quite constrained in comparison to other countries. Platforms that work with data may avoid the use of data for commercial purposes and focus on offering a service that allows people to share data with others on third-party platforms. Even in these constrained environments, some questions arise: Are platforms changing their behavior because of the data? Is data changing the pricing policies of platforms? To what extent is data influencing platforms and customers?

Highlight 9. Digital platforms are new and innovative products. But many insights about compatibility between devices still apply to digital platforms. When the focus of compatibility is not the the commercial use of data but the number of users who use the platform, all the traditional insights about compatibility still apply. Independently of what approach we consider, compatibility tends to be commoner among end-platforms, and it tends to create incentives to increase prices. However, when the commercial use of data is involved, this may not be true.

[21] See Farrell and Saloner (1985) for an analysis of the problem of adopting standards and the role of communications.

[22] See Katz and Shapiro (1986) for an analysis of standards.

[23] https://www.thesun.co.uk/news/2067929/uber-london-banned-tfl-petition-ceo-sadiq-khan/

2. Data. The New Compatibility

2.1. The Vertical Relationships

In Section III.1, we pointed out that we can find two types of data relationships among digital platforms. The first one was the "vertical one". Platforms that provide others with data to produce something. This behavior is common among platforms that do not compete for the same set of customers directly. It may seem that this case is unrelated to compatibility. However, platforms are built upon other platforms, and some of them communicate with other platforms using compatible communication protocols. This compatibility is well-known among engineers. However, among economists it is less noticeable because standards, adapters or similar products or devices are not so visible as they were in the 80s and 90s. However, they play a relevant role today, and because of those standards and compatible protocols, data generated by some platforms can be used by other platforms. The problem is how that data is used, and if that data may generate inefficiencies (such as the double-marginalization).

Let us illustrate this case. Susan loves to eat healthy food. Normally, she orders from different apps when she finds a good offer. However, the company who owns the app has paid large fees to Facebook and Google to know the habits of people like Susan. Obviously, those fees are costs for the app company, and it has to charge a bigger price in each order to cover those costs. The inefficiencies arise because the "app company" does not take into account that the platforms are charging a price with a markup when they sell the data.[24]

On the other hand, companies that are able to integrate the extraction of data within the platform will not create this inefficiency. For example, if Susan uses Amazon or UberEats. It is possible that big platforms with a lot of users are not buying data to other companies. In that sense, they will not have to charge higher prices. In fact, maybe, they set even lower prices. However, these inefficiencies are not only related to data. They can appear in other digital markets, such as video-consoles. For example, Susan also loves to play video-games in her video-console.[25] In the video-console market, there are platforms (video consoles) that are used to play video games (users, first side) created by developers (second side). Clearly, both the platform and the developers want

[24] More formal: if the upstream market operates as an oligopoly, the firms' equilibrium prices contain a markup, which downstream firms treat as part of their marginal costs. This creates the inefficiency.

[25] An outstanding work that addresses the video console market from an empirical point of view is Lee (2013). If you prefer a theoretical approach, see Hagiu (2004). For an introduction to this topic, see Evans, Hagiu and Schmalensee (2008).

to earn profits. In the video console market, platforms sell the hardware below costs and make profits from selling video games. Developers pay a fee for developing titles, and users pay for playing the game. In this scheme, it is clear that a double marginalization problem is possible as long as developers take an input (Developers' toolkit) and produce a game with it.[26]

We have claimed that companies that can capture data themselves can offer services at lower prices. However, this is only true for those platforms that sell products in which data is relevant. If we turn back to Figure 1, we mainly refer to the end and middle platforms. This is the reason behind the integration of killer apps or the acquisition of killer apps by big players that operate infrastructural or middle platforms. For example, killer apps tend to be integrated by the upstream platforms. Examples of this behavior are WhatsApp (integrated with Facebook) or Paypal (integrated with eBay). There are several reasons why companies integrate those killer apps: because they can damage other platform's products because those apps do not take into account their effects on other platforms (see Viecens, 2009), or because it is more profitable for the platform (see Economides and Katsamakas, 2006). Even policymakers would be interested if they could increase welfare (see Nocke, Peitz and Stahl, 2007). For example, WhatsApp used to charge an annual fee. However, in 2016, after its purchase by Facebook in 2014, it became free. When WhatsApp was not integrated, its prices were inefficiently high. Once it was acquired, the integration led to a zero price. The double-marginalization problem was solved.[27]

Nonetheless, the inefficiencies that foster integration get weaker and tend to disappear when substitutability among the applications is high.[28] This clearly resembles the case of instant messaging services. Currently, many services co-exist with a high degree of substitutability (Telegram, Line, WhatsApp, etc.). All of them are free to use, but none of them are compatible. This is an example of how substitutability have lowered the inefficient high prices that were the consequence of the double-marginalization issue.[29]

---

[26] Other services susceptible of having these inefficiencies are the video-streaming services (HBO, Netflix, Hulu, etc.). These services operate in a similar way that the cable TV, which was pointed out as a market with double-marginalization. See Waterman and Weiss (1996).

[27] Making WhatsApp free was the strategy of Facebook to make customers pay for other services. https://techcrunch.com/2014/02/24/whatsapp-is-actually-worth-more-than-19b-says-facebooks-zuckerberg/

[28] Integration is not always the best option. Hagiu and Wright (2015) prove that optimal integration depends on the market structure. There is always a trade-off between integration or disintegration.

[29] This price reduction as the consequence of the integration is not exclusive of digital platforms. Economides and Salop (1992) also point out that the integration of the complementary companies reduces the total price of the complementary goods. However, Viecens (2009) proved that, in digital platforms, integration and substitutability mitigate the double-marginalization problem.

Also, in the smartphone ecosystem, apparently, there is no double-marginalization problem. The main reason is that, either infrastructural platforms are open source, such as Android; or they are vertically integrated, such as iOS. Nonetheless, in Figure 1, we observe that data is generated by final users, and that data is losing relevance when we move towards the infrastructural platforms. Normally, this problem may arise between the middle and end platforms, but it is not clear to what extent it is a generalized phenomenon. Some middle-platforms have private or open Application Programming Interfaces (APIs) that can be used by developers to create new services. Some companies (such as Garmin) prefer to sell the access to developers, but others (such as Runtastic) prefer to give it for free. It depends on the business model of each company, and the strategy followed. Some companies prefer to give it for free to boost the creation of an ecosystem, others prefer to sell the API to monetize the data. However, these pricing decisions may change over time. For example, Garmin or Under Armour APIs were free some years ago, but right now accessing those APIs requires a payment. To what extent there is a double-marginalization in these cases is unknown.

In other cases, data are used within platforms to help developers foster the adoption of their apps to increase the relevance of the ecosystem. For example, the platform Steam developed by Valve[30] allows users to have a digital library with all their games available worldwide. The platform is free for users, and developers only pay for developing games. Nonetheless, both, users and developers, generate a tremendous amount of data. This data is not only helping Steam to know which games are the most played but also, to gather information about the users' hardware, their willingness to pay for games, which genres are more interesting, etc. All that information is used to help developers find their place in the market.[31]

All those cases illustrate how platforms behave with regard to data in a vertical sense. As a summary, in this case, data is not creating new issues. Data is only a new input (a very valuable one), but the intuitions are not changing radically. Although traditional insights remain valid (see Weyl, 2008 or Viecens, 2009), this statement does not imply that the analysis has to be the same than with traditional markets.[32]

Lastly, let's re-take the case of Susan. While Susan was running, different digital platforms were taking different types of information (Spotify and

[30] http://store.steampowered.com/

[31] https://partner.steamgames.com/doc/marketing

[32] In empirical terms, this creates another issue: the market identification. In other words, how to know where are the market boundaries. Market identification is beyond the scope of this chapter. See Filistrucchi, Geradin and Van Damme (2012).

Facebook were taking information about the songs played, and the friends nearby respectively). Others were taking the same type of information (Garmin and Google Maps were tracking her GPS position, but probably for different purposes). Lastly, some platforms were sharing their data with competing platforms. For example, Garmin was extracting data from the Garmin device, and MyFitnessPal was synchronizing the information of the device with its online platform.

This case is the most interesting one. She is using a Garmin device that, automatically, synchronizes with Garmin Connect (the digital platform of Garmin). However, MyFitnessPal allows her to synchronize Garmin data with MyFitnessPal. This behavior is totally new. Why does a company allow its competitors to access its information?

Highlight 10. Compatibility between vertical companies (provider-client) highlights the role of data as a mere input that is created by some companies and exploited by others in a different market. In this case, vertical integration between those companies may lead to lower prices, but it depends on how relevant is the double-marginalization. Nonetheless, the incentives for integrating tend to disappear when substitutability among companies' products is high.

2.2. The Horizontal Relationships

In Section III.1, we point out that some platforms may share their data with their competitors, and in previous sections, we have introduced this case, but it was incomplete, and we only pointed out some examples, such as the fitness tracker market, in which several platforms allow their competitors to access their data. Let us focus on this case.

Let us return to Susan. Susan was using a Garmin device and the MyFitnessPal app. These are two competing companies. Garmin owns a digital platform (Garmin Connect) in which all the data of their wearables is synchronized. On the other hand, MyFitnessPal is a digital platform, but it is provided by Under Armour, which has its own devices too. In this case, advertisers or developers who want to access MyFitnessPal data (to promote a product or to develop a new app) will find that not all users are equal. Some of them are pure Under Armour users, but others are users of Garmin, Fitbit, etc.[33] In comparison with previous sections, in this case, users are not accessing to a bigger pool of users

[33] Some consumers may use their smartphones to workout, many of them have GPS, accelerometer, etc. so, they may be used them as a fitness device. However, for simplicity's sake and without loss of generality, we omit this case because the main purpose of a smartphone is not the fitness tracking.

(like in the Katz and Shapiro's definition). Instead, advertisers, developers, sellers, etc. are the ones who access to a bigger pool of users.

Traditionally, it was thought that compatibility could increase (see Farrell and Saloner, 1985; Katz and Shapiro, 1985), or decrease (see Matutes and Regibeau, 1988) price competition. Some authors argue that the net effect of compatibility in prices was influenced by the product diversity, the total output, the users' valuation of the whole system, etc. However, the current evidence points out that, in digital platforms, sharing networks or databases mitigate the price competition among platforms.[34]

Nonetheless, these changes in prices are not easy to notice. Evaluating the prices of digital devices such as wearables is not easy. Digital platforms are influenced by the competition with other producers, technological change, network compatibility, market segmentation, etc. For example, in the wearable market, platforms invest a lot of money in R& D to outperform their competitors. This behavior starts a "quality race".[35] Technological change imposes a challenge to those who want to study prices because almost every year a new generation is launched, and during the year, new products are launched that compete with the incumbents. All those changes make quite difficult to test if compatibility is increasing or reducing prices in a specific market.

However, if we only pay attention to the compatibility, and we omit for a moment the technological, other effects appear. There is an incentive to increase prices in platforms as a consequence of compatibility that is exclusive of platforms. Compatibility mitigates the incentives to reduce prices to attract some customers. In fact, in the fitness-tracker market, we observe this pattern. Although in the next section we will pay attention to it, in this section, let us focus on why companies allow other competing platforms to access their own database.

Let us consider a fitness-tracker company such as Fitbit or Garmin. Currently, they sell a device with an integrated digital platform. The digital platform attracts a lot of users, but to attract more users these companies need more functionalities, more apps, and better interfaces. To do so, they need to attract developers too. In this situation, they can decide to reduce users' prices. With this policy, platforms want to attract a lot of users interested in the device and the platform. This price reduction increases the users' base, and at the same time, many developers start to be interested in developing applications for the platforms. Companies have given up market power and profits in the users'

[34] See Doganoglu and Wright (2006) and Salim (2009). See Sánchez-Cartas and León (2017) for a generalized model.

[35] See Salim (2009). She develops a model in which quality races are endogenously generated.

side to boost the adoption of their products and the overall profits.[36] This is the intuition of a two-sided business model.

However, some companies have realized that some competitors' networks have open APIs to attract developers. These APIs allow others to access and export information about users. In this situation, many companies have created an extra functionality that allows users in those platforms to "migrate" to other platforms automatically. In some cases, all the information about users can be synchronized in several platforms, and many users are interested in doing that because some platforms offer extra information about the calories burnt, performance, etc. that the other platforms cannot. That implies that some users synchronize their data with other companies' platforms even when they have not bought the companies' device. This practice allows companies to relax their policy of low prices for devices. Users are coming into the platforms from competitors'. Developers are happy because the users' base is increasing, and there is no reason to keep low prices for the devices.[37] This example illustrates a case that resembles what is going on in the fitness-tracker market. The possibility of accessing the users' data in other platforms reduces the interest of platforms in subsidizing their devices to attract consumers.[38] In comparison with incompatible digital platforms, compatibility increases the market power of platforms because they relax their competition. The network effects between the sides lose relevance. Nonetheless, it is possible that some users will use different platforms at the same time (multihoming). If users can easily use two platforms at the same time, the incentives to become compatible disappear. However, multihoming is not always a good substitute for compatibility (see Doganoglu and Wright, 2006), especially for users, who have to pay for using two platforms that do not allow them to export their data. Compatibility and

Highlight 11. Compatibility among competitors leads to higher prices on at least one side of the market. It mitigates the incentive to reduce prices to attract more consumers because the network is shared with competitors. However, we have to take care of not confusing compatibility with multihoming. Compatibility implies being on a platform and being able to access others from that platform. Multihoming implies being on several platforms at the same time.

[36] Developers will be willing to pay more to access your huge database, so you expect larger profits.

[37] For a technical explanation see Doganoglu and Wright (2006), Salim (2009) and Sánchez-Cartas and León (2017).

[38] This argument can be stated the other way around. The compatibility may reduce the incentive to subside developers to attract users because users can connect with anybody on another platform.

multihoming mitigate competition and increase profits, but it is not clear which one is preferred over the other (especially, in terms of welfare).

2.3. Do Policymakers Have to Worry about Compatibility?

The previous examples raise a clear concern about the use of data by platforms. Apart from the already known issues about privacy in digital services, a new front is open. In previous sections, we have argued that companies may have an incentive to share data that can increase the consumers' prices. In this sense, it seems that consumer welfare will be damaged by this practice. However, the problem is not so simple. Let us consider the users only. Compatibility among platforms may have a clear advantage for users, who may export their data to the platform they prefer without reducing the number of platforms in the market. Other users may also benefit from the possibility of using combinations of wearables or devices such as a smart balance of Withings and the fitness tracker of Fitbit. Compatibility may also increase the incentives to compete. Higher compatibility implies that is easier to compare platforms, so they can be forced to produce platforms with more quality or more functionalities at the same price. Obviously, if we omit those benefits that arise from linking the platforms, it seems that compatibility may harm users. Those users who buy the device and do not care about which platform they use will be harmed by this policy. They would pay a higher price because of the compatibility. However, in terms of welfare, it is not clear which group is more numerous nor the net change in welfare.

Nonetheless, customers of digital platforms are not only users, developers are also customers.

To measure the impact of compatibility, we need to take them into account. In this sense, it is clear that developers benefit from the compatibility in different ways. The most obvious one is the possibility of accessing a large pool of users, but it is not the only one. Compatibility among platforms also reduces the number of protocols and complexity of databases. Having a common way to communicate among services allows developers to work more efficiently in different frameworks. However, it is also true that they may pay a higher price.

As a summary, from a strict point of view, it is not clear if compatibility increases or reduces welfare,[39] there are forces in both directions. On the other hand, the increase in market power of platforms as a consequence of

[39] There is theoretical evidence in this sense, Salim (2009) proves that compatibility is welfare enhancing, but her model does not cover all the potential scenarios.

compatibility may not be superior to the market power of a company that sells a device that is not influenced by network effects, such as watches or clothes. Intervention from public authorities may not be justified in this case.

However, it is clear that some digital platforms have a dominant position in the market (Google, Amazon or Facebook). However, compatibility can be rejected as the driver of these monopolies. In most of the cases, it is the own nature of digital platforms and their network effects which motivate a situation of dominance. Compatibility can help in increasing this dominance, but it is not the main driver (see Sánchez-Cartas and León, forthcoming). One clear example is the fitness tracker market, in which several companies compete, and there is no clear dominance.

Nonetheless, it is true that compatibility may create perverse incentives in markets in which "the-winner-takes-all-the-market" outcome is a possible result. In these markets, small players may be interested in sharing their databases with a leader because, in that way, the differentiation between them and the leader would be larger. In this case, they could create two different markets, one for data and another one for devices. For example, let's imagine a wearable market in which there are two companies: the leader and the follower. The leader has a bigger network as a consequence of being an incumbent in digital markets, and it sells average-quality devices. The follower has a tiny network, but it sells high-quality devices. Both of them sell a device to users and a platform to developers. However, because of the network effects, the leader has a clear advantage, and it can almost expel the follower from the market. The follower has a great device, but without a powerful platform, its growing capacity is limited. If the follower agrees to share its data with the leader, that increases the size of the leader network, and the leader can focus on the platform. On the other hand, because of the compatibility, the follower can focus on the device and monopolize the market of devices. Both companies benefit as long as the monopoly profits of the two markets are higher than the profits in the initial situation of duopoly.

This is a fictional scenario, and it is not clear how likely it is. Nonetheless, competition authorities may consider this possibility in new markets, such as the Internet of Things markets. In these markets, some companies can focus on selling devices only if there is a great pool of users who only care about the device itself (and not about the communities or the linked services). This phenomenon is already common in the fitness tracker market, where there are users who only value the device and do not care about working out with other people nor sharing their performance with others.

Lastly, a point worth emphasizing is that compatibility is not a type of merger or tacit collusion. With compatibility, there is no coordination in the decision-making process of companies. In contrast with mergers, agreements are not required. Neither they are needed to behave in the same way, as we expect when there is tacit collusion. Compatibility can be asymmetric and, in many cases, it is asymmetric, and it can arise from the desire of only one company (if there are no legal barriers to it). Nonetheless, some platforms may cooperate when they become compatible, for example, to develop new technologies. This cooperation may lead to markets where there is tacit collusion (one platform becomes a high-quality vendor, and the other one a low-cost one). Even in this case, it is not clear if the welfare will increase or decrease.[40]

Highlight 12. A priori it is not clear the impact on the welfare of the compatibility among competitors. Even without considering the profits of the platform, it is not clear whether or not all sides benefit. It will depend on each market. Nonetheless, there is no reason to think that compatibility will lead to the "winner-takes-all-the-market" outcome. However, it is true that the companies involved in those compatibility agreements increase their market power.

3. An Example of Digital Platform Market: The Fitness Tracker-market

In the previous sections, we have been using the fitness tracker as an example. In this section, we focus on this market to show the relevance of compatibility. However, an extensive analysis of the market is beyond the scope of this work. This market involves a smart device (a wearable), and a digital platform that links the device with other smart devices such as tablets or smartphones. The first two companies of this market in achieving notoriety were Fitbit and Jawbone in 2011.[41] They started by selling a device. The platform idea came later on when they, and other competitors, realized that it was time to attract more users by creating communities,[42] and developers by creating larger platforms and ecosystems.[43]

---

[40] In fact, Salim (2009) points out that *cooperative investment by standardized platforms might create higher aggregate surplus than [non-compatible platforms].*

[41] http://www.businessinsider.com/the-smartwatch-and-fitness-band-market-2015-1

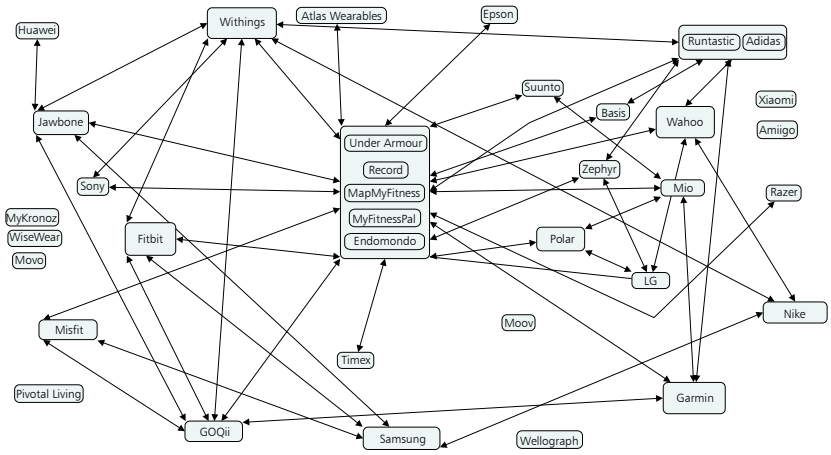
[42] The linking with social networks took place in 2012. http://mashable.com/2012/01/27/ facebook-privacy-open-graph/#uv_7foC0jsqY

[43] It is easier to convince people to buy a high-quality product than to convince them to buy a product that will be high-quality only when other consumers adopt it. And it is easier to convince similar people than to convince heterogeneous people to adopt the product.

Once platforms were established, the market started to grow. New platforms entered the market, and users started asking for more functionalities. Then, opening the network to competitors was slowly taking place as a way to keep the users who wanted to have functionalities of different platforms. In Figure 3 a network that represents the compatibility relationships among the databases of the relevant players in the fitness tracker ecosystem in July 2016 is depicted.[44] The most connected player is Under Armour. The professional access to their API is not free. However, some years ago, it was free. Garmin is another example of this behavior. They have a one-time license fee of $5000, although until 2014 it was completely free. However, other companies have open APIs because: a) a fitness-tracker is not the main line of business (as Nokia-Withings), or b) their ecosystems are not so vibrant as those of Garmin or Under Armour.[45] However, what is truly interesting about the Figure 3 is the complex network of relationships among the platforms. Obviously, many users take advantage of this compatibility, but probably other multihome. Nonetheless, compatibility

FIGURE 3

RELATIONSHIPS AMONG DATABASES OF FITNESS TRACKER COMPANIES. SUMMER 2016



[44] We only consider those companies which sell a fitness tracker. There are other players that influence the market such as Google Fit, Apple Health or Runkeeper, but they do not sell a fitness tracker with a complementary platform.

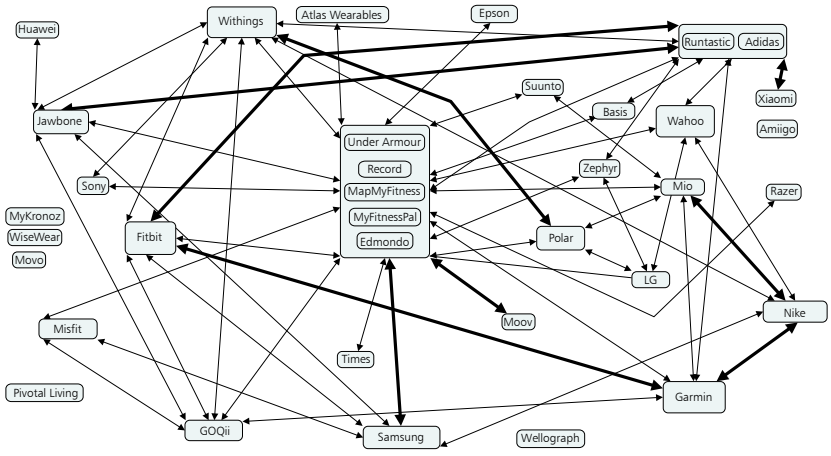
[45] Under Armour: https://developer.underarmour.com/, Garmin: https://goo.gl/nLUw35 and https://goo.gl/QkHfHu

is much easier to notice than multihoming. In Figure 4, we can observe the situation of the fitness tracker ecosystem in June 2017. The bold lines represent the new connections that have appeared between July 2016 and July 2017. We observe that nine new connections have appeared. This change in only one year highlights how relevant is the compatibility issue for companies in this market.

On the other hand, we have stated that some companies may behave as multi-sided platforms or maybe as sellers of devices or platforms. In this sense, it is interesting to consider the Terms & Conditions (TOC) of the service provided by the fitness tracker platforms. In Table 1, we observe a list with all the relevant players in the fitness tracker market, the link to their TOC, the last update of the TOC, and information about their behavior towards the users' data.[46] Obviously, the degree of sharing differs from company to company. We only highlight those who are willing to share some non-personal information with third-party companies. In some cases, companies state that they offer the possibility of connecting to third-party networks, but the transfer of data is up to the user. From a privacy point of view, this is a clear disclaimer. But the interesting point is that, at this moment, a common pattern in the industry is to allow the sharing of non-personal data.

FIGURE 4

RELATIONSHIPS AMONG DATABASES OF FITNESS TRACKER COMPANIES. SUMMER 2017



[46] This data was cross-checked in November 2017.

TABLE 1

TERMS AND CONDITIONS. PRIVACY POLICIES

Company	*Date*	*Is sharing data allowed?*	*Link*
Under Armour	22/01/2016	Yes	https://goo.gl/eDwUig
Jawbone	16/12/2014	No	https://goo.gl/aYZ6qv
Mio	No date	Yes	https://goo.gl/KES6b1
Suunto	No date	Yes	https://goo.gl/4ENzLh
Garmin	14/02/2017	Yes	https://goo.gl/ttnsBG
Fitbit	28/07/2016	Yes	https://goo.gl/JGGdt4
Mi (Xiaomi)	06/05/2016	Yes	https://goo.gl/1mtKZd
Apple	19/07/2017	Yes	https://goo.gl/x2joJg
Withings	20/07/2017	Yes	https://goo.gl/S14zoq
Mykronoz	20/10/2015	Yes	https://goo.gl/a397KA
Huawei	01/07/2016	Yes	https://goo.gl/iPbo8r
Epson	01/02/2012	No	https://goo.gl/gkFXms
Wisewear	01/01/2015	No	https://goo.gl/BuRbXB
Atlas	No date	Yes	https://goo.gl/9PqTiD
Amiigo	Out-of-business		
Razer	01/04/2016	Yes	https://goo.gl/uefz14
Wellograph	04/04/2014	No	https://goo.gl/Q71Thg
Runstastic (Adidas)	13/04/2017	Yes	https://goo.gl/cXTr8P
Misfit	No date	Yes	https://goo.gl/XqGKi9
Wahoo	No date	Yes	https://goo.gl/bxsRFy
GOQii	07/04/2017	Yes	https://goo.gl/aUNgzf
Samsung	22/03/2016	Yes	https://goo.gl/HjLSUa
Basis (Intel)	No date	Yes	https://goo.gl/VeP9B6
Polar	No date	Yes	https://goo.gl/nKJUri
Sony	01/04/2015	Yes	https://goo.gl/ny5tu1
Zephyr	01/06/2016	Yes	https://goo.gl/SqHcTv
Timex	27/04/2015	No	https://goo.gl/T6DrMD
Moov	01/08/2016	Yes	https://goo.gl/ZvaQWm
Adidas	No date	No	https://goo.gl/zUNnXS
Pivotal Living	Out-of-business		
LG	No date	Yes	https://goo.gl/eCgD1v

V. CONCLUSIONS. THE CHANGES THAT DIGITAL PLATFORMS HAVE BROUGHT

Digital platforms are everywhere. In our daily life, we can use dozens of them without noticing it, but they are impacting our lives, and they are growing in relevance. These digital platforms are also bringing new behaviors

and challenging our knowledge of how companies compete. In this work, we analyze the relationships of platforms with each other. We focus on those relationships in which there are exchanges of data. These relationships are the most interesting ones because to transfer data, platforms must be compatible with each other, either in communication protocols or in data formats. Then, we analyze different classification of platforms depending on their relationships. We differentiate between vertical relationships and horizontal ones. The vertical relationships represent the idea that some platforms depend on others to work but also, they represent the relationships of platforms that sell data to other platforms in different markets. On the other hand, the horizontal relationships represent an exclusive feature of digital platform markets: the exchanges of data among competitors. We analyze these two classifications following different examples of real digital markets, and we show how different economic intuitions are still valid in digital platform markets. Nonetheless, we pay special attention to those cases in which new intuitions emerge. In this work, we also show how data can play different roles in markets depending on who uses the data and who is providing that data. We focus on the role of those relationships with data from a regulator's point of view, and we highlight that it is not clear if those exchanges of data are increasing or reducing welfare.

Lastly, we focus on a real case: The fitness tracker market. This market is characterized by a lot of exchanges of data among competitors. We depict the current network of relationships among the most relevant companies in the market and how that network has evolved. We also analyze the terms and conditions of use of those companies, and we show that the vast majority of them are open to sharing data with third-party companies.

BIBLIOGRAPHY

Affeldt, P. L. (2011), "Tying and bundling in two-sided markets," Tilburg University Master Thesis, Department of Economics.

Affeldt, P.; Filistrucchi, L., and T. J. Klein (2013), "Upward pricing pressure in two-sided markets," *The Economic Journal,* 123: 505–523.

Ambrus, A., and R. Argenziano (2004), "Network markets and consumer coordination," *CESifo Working Paper,* 1317.

Anderson, S. P., and S. Coate (2005), "Market provision of broadcasting: A welfare analysis," *The Review of Economic studies,* 72: 947–972.

Armstrong, M., and J. Wright (2007), "Two-sided markets, competitive bottlenecks and exclusive contracts," *Economic Theory,* 32: 353–380.

Belleflamme, P., and M. Peitz (2015), *Industrial organization: Markets and strategies,* Cambridge University Press.

Belleflamme, P., and E. Toulemonde (2004), "Emergence and entry of b2b marketplaces," *CORE Discussion.*

Bolt, W., and A. F. Tieman (2005), "Skewed pricing in two-sided markets: An io approach," Technical report, *DNB Working Paper,* 13.

Caillaud, B., and B. Jullien (2001), "Competing cybermediaries," *European Economic Review,* 45: 797–808.

— (2003), "Chicken & egg: Competition among intermediation service providers," *RAND Journal of Economics:* 309–328.

Doganoglu, T., and J. Wright (2006), "Multihoming and compatibility," *International Journal of Industrial Organization,* 24: 45–67.

Economides, N. (1996), "The economics of networks," *International Journal of Industrial Organization,* 14: 673–699.

Economides, N., and E. Katsamakas (2006), "Two-sided competition of proprietary vs. open source technology platforms and the implications for the software industry," *Management Science,* 52: 1057–1071.

Economides, N., and S. C. Salop (1992), "Competition and integration among complements, and network market structure," *The Journal of Industrial Economics:* 105–123.

Evans, D. S. (2003), "Some empirical aspects of multi-sided platform industries," *Review of Network Economics,* 2,3:191-209

— (2011), *Platform economics: Essays on multi-sided businesses,* Competition Policy International.

Evans, D. S.; Hagiu, A., and R. Schmalensee (2008), *Invisible engines: How software platforms drive innovation and transform industries,* MIT press.

Evans, D. S., and R. Schmalensee (2005), "The industrial organization of markets with two-sided platforms," *Technical report,* National Bureau of Economic Research, Cambridge, MA.

— (2013), "The antitrust analysis of multi-sided platform businesses," *Technical report,* National Bureau of Economic Research, Cambridge, MA.

Farrell, J., and G. Saloner (1985), "Standardization, compatibility, and innovation," *The RAND Journal of Economics:* 70–83.

Filistrucchi, L. (2008), "A ssnip test for two-sided markets: The case of media," *NET Institute Working Paper,* available at SSRN 1287442.

Filistrucchi, L.; Geradin, D., and E. Van Damme (2012), "Identifying two-sided markets," Dipartimento di Scienze Economiche, Universit`a degli Studi di Firenze, *Working Paper,* 01/2012.

Filistrucchi, L., and T. J. Klein (2013), "Price competition in two-sided markets with heterogeneous consumers and network effects," *NET Institute Working Paper,* available at SSRN 2336411.

Hagiu, A. (2004), "Two-sided platforms: Pricing and social efficiency," *Technical report,* Research Institute of Economy, Trade and Industry (RIETI), Japan.

Hagiu, A., and H. Ha Laburda (2014), "Information and two-sided platform profits," *International Journal of Industrial Organization,* 34: 25–35.

Hagiu, A., and J. Wright (2015), "Multi-sided platforms," *International Journal of Industrial Organization,* 43: 162–174.

Ivaldi, M.; Sokullu, S., and T. Toru (2011), "Airport prices in a two sided framework: an empirical analysis," in *CSIO/IDEI 10th Joint Workshop on Industrial Organization,* Toulouse.

Jullien, B. (2005), "Two-sided markets and electronic intermediaries," *CESifo Economic Studies,* 51: 233–260.

Kaiser, U., and J. Wright (2006), "Price structure in two-sided markets: Evidence from the magazine industry," *International Journal of Industrial Organization,* 24: 1–28.

Katz, M. L., and C. Shapiro (1985), "Network externalities, competition, and compatibility," *The American Economic Review,* 75: 424–440.

— (1986), "Technology adoption in the presence of network externalities," *Journal of political economy,* 94: 822–841.

Lee, R. S. (2013), "Vertical integration and exclusivity in platform and two-sided markets," *The American Economic Review,* 103: 2960–3000.

Matutes, C., and P. Regibeau (1988), ""mix and match": Product compatibility without network externalities," *The RAND Journal of Economics:* 221–234.

Nocke, V.; Peitz, M., and K. Stahl (2007), "Platform ownership," *Journal of the European Economic Association,* 5: 1130–1160.

OECD (2009), "Two-sided markets," *Technical report,* OECD. Policy Roundtables, Competition Law and Policy, Paris.

Parker, G. G., and M. W. Van Alstyne (2000), "Internetwork externalities and free information goods," in *Proceedings of the 2nd ACM Conference on Electronic Commerce:* 107–116, ACM.

— (2005), "Two-sided network effects: A theory of information product design," *Management science,* 51: 1494–1504.

Rochet, J.-C., and J. Tirole (2003), "Platform competition in two-sided markets," *Journal of the European economic association,* 1: 990–1029.

— (2006), "Two-sided markets: a progress report," *The RAND journal of economics,* 37: 645–667.

Rysman, M. (2009), "The economics of two-sided markets," *The Journal of Economic Perspectives,* 23: 125–143.

Salim, C. (2009), "Platform standards, collusion and quality incentives," *Discussion Paper,* 257, Governance and the Efficiency of Economics Systems, Free University of Berlin.

Sánchez-Cartas, J. M., and G. León (forthcoming), "Assessing compatibility and competition issues in wearable markets," *International Journal of Economic Theory.*

— (2017), "Shared networks and market power in two-sided markets," *Economics Bulletin,* 37: 2173–2180.

— (2018): Multi-sided markets: A literature review, *CAIT Working Papers.*

Schiff, A. (2003), "Open and closed systems of two-sided networks," *Information Economics and Policy,* 15: 425–442.

Song, M. (2013), "Estimating platform market power in two-sided markets with an application to magazine advertising," *Simon School Working Paper,* no FR 11-22.

Tiwana, A. (2013), *Platform ecosystems: aligning architecture, governance, and strategy,* Newnes.

Van Damme, E.; Filistrucchi, L.; Gerardin, D.; Keunen, S.; Klein, T. J.; Michielsen, T., and J. Wileur (2010), *Mergers in two-sided markets: A report to the nma,* Dutch Competition Authority.

Viecens, M. F. (2009), "Pricing strategies in two-sided platforms: the role of sellers' competition," *Documento de trabajo,* 11.

Waterman, D., and A. A. Weiss (1996), "The effects of vertical integration between cable television systems and pay cable networks," *Journal of Econometrics,* 72: 357–395.

Weisman, D. L., and R. B. Kulick (2010), "Price discrimination, two-sided markets, and net neutrality regulation," *Tul. J. Tech. & Intell. Prop.,* 13: 81.

Weyl, E. G. (2008), "Double marginalization in two-sided markets," *Working Paper.*

— (2010), "A price theory of multi-sided platforms," *The American Economic Review,* 100: 1642–1672.

Wright, J. (2004), "One-sided logic in two-sided markets," *Review of Network Economics,* 3.

ASYMMETRIC INFORMATION AND REVIEW SYSTEMS: THE CHALLENGE OF DIGITAL PLATFORMS[1]

Michelangelo ROSSI

Abstract

In this chapter we review theoretical and empirical works related to the issues of asymmetric information and the role of review systems in digital contexts. First, the concepts of Adverse Selection and Moral Hazard are introduced as they form the two main classes of issues related to the asymmetries of information between parties. Later, we describe the common design of review systems and discuss the empirical evidence of the impact of reviews on the performance of online users. Finally, since feedback systems can simultaneously reduce Adverse Selection and discipline Moral Hazard, we clarify the signaling and the sanctioning roles of reviews describing the theoretical mechanisms behind these functions; and the empirical findings from several digital marketplaces.

Key words: Digital platforms, asymmetric information, moral hazard, adverse selection.

JEL classification: L14, L15, L86.

[1] I thank the conferences participants at the 5th Madrid Microeconomics Graduate Workshop for useful feedback. I am grateful to my advisors, Natalia Fabra and Matilde Machado, for their patient guidance and advice. Special gratitude to Elizaveta Pronkina for her enormous support and encouragement. All errors are mine.

I. INTRODUCTION

Digital platforms such as eBay, Amazon or Airbnb have achieved enormous success and popularity in the last two decades and they keep attracting new clients. Now, online marketplaces connect millions of people around the world and digital commerce exerts a significant impact on the GDP growth of many countries.[2]

Interestingly, the growth and the expansion of digital commerce was underestimated by many economists a few years ago: what is now a customary habit for millions of users was taken with surprise and skepticism. In particular, some characteristics of online trade such as anonymity were considered an insurmountable limit that would have prevented the formation of the trust among parts, essential for transactions. To understand the skeptical attitude towards online transactions, it is worth to recall the story of one of the very first items sold on eBay (at that time called AuctionWeb): a broken laser pointer. In 1995, a few months after the website launch, the eBay founder Pierre Omidyar decided to sell online his broken laser pointer; in the listing description, he wrote that the item was indeed damaged. Still, after a few weeks the pointer was sold for 14.83 US dollars. Surprised by the final price, Omidyar contacted the buyer asking whether it was clear to him that the laser pointer was broken. The buyer responded he was a "collector of broken laser pointers".[3]

This anecdote is often cited to remark the limitless variety of buyers and sellers that can be matched through online platforms. However, it is important to note that, at that time, even the eBay founder casted some doubts on the success of online anonymous transactions. In his question to the winning bidder, he implicitly pointed out one of the issues that could potentially hinder exchanges in digital platforms.

First, the two sides of online transactions do not have access to the same pieces of information about the object of the transaction: for instance, eBay sellers are usually much more aware of the quality of the items they are selling relative to potential buyers; in the same way, Airbnb hosts have a better understanding of the location of their dwellings with respect to the guests who are going to rent their apartments.

Moreover, the two sides can partially determine the transactions' quality through their actions: in eBay, sellers choose how to organize the delivery

[2] McKinsey Global Institute reports show that Internet accounted for more than 20 percent of GDP growth in developed countries over the last five years (Manyika and Roxburgh, 2011; Manyika, *et al.,* 2016).

[3] This and other stories about the eBay early years can be found in Cohen (2003).

process for the listed objects; similarly, in Airbnb, hosts can decide how much effort to put in cleaning their apartments; and guests may respect or not the house's rules.

In this sense, we can define two potential issues of online transactions related to the anonymity and the distance among users:

- The two sides involved in the online transactions have different levels of information regarding the inner quality of the service, that can hardly be modified by users' action. In the most extreme cases, one side (typically the buyer side) is aware of the service's quality only after the transaction has occurred. Because of this, the price that the least informed side is willing to pay for the transaction will take into account the quality uncertainty and it will reflect an "average" expected quality level. Accordingly, the sellers with high quality will be driven out of the market by the low prices and, using economic jargon, the sellers will be adversely selected (as only the ones with low quality are willing to be on the market). In the remaining part of the chapter, we will call this potential issue as *Adverse Selection.*

- The quality of the transactions depends on the level of attention, effort and care that the two sides put in the process. Still, the transaction price is often decided before the effort choice is made and the two parties may be tempted to not accomplish their duties after the money transfer. Such behavior could be indeed very likely in online markets since users seldom interact with each other more than once and their misbehavior cannot be punished in future periods. All this leads to another type of uncertainty regarding the services' quality. We will denote it as *Moral Hazard.*

These two issues are potentially present simultaneously in all digital platforms; however, the dominance of one over the other depends on the capacity of one side to vary the quality of the service with his actions. For example, we may expect to observe the prevalence of Adverse Selection issues in platforms where the quality depends less on the effort decision such as Booking or Expedia. Differently, Moral Hazard may turn out to be dominant in platforms like Uber or BlaBlaCar since the drivers' performance directly defines the quality of the service provided.

Despite these weaknesses, several digital platforms found their path to success and online trade is under enormous growth. Part of this success may depend on the way digital platforms tried to reduce these issues with an innovative solution: review systems. First introduced by eBay, almost all the digital

platforms implemented feedback systems thanks to which users can review their experiences in previous transactions. Reviews reduce Adverse Selection issues since new pieces of information increase the precision of the buyers' estimate about seller quality; besides, they also mitigate Moral Hazard issues and the history of past reviews creates a reputation regarding the users' on-going behavior that can lead to potential punishment after some misconducts. In this sense, review systems play at the same time the role of a signaling and sanctioning device, notions firstly introduced by Dellarocas (2006).

In this chapter we are going to describe how review systems work; and, in particular, how they discipline the Adverse Selection and the Moral Hazard issues in digital platforms.

The chapter consists of five parts: in the next part we analyze the impact of reviews over the performance of users in different platforms with a focus on the main drawbacks of review systems. In parts three and four we discuss Adverse Selection and Moral Hazard separately. Part five presents recent works about the joint impact of review systems on Adverse Selection and Moral Hazard, and how these two issues are connected. Part six concludes the chapter.

II. REVIEW SYSTEMS: DESCRIPTION AND IMPACT

In the introduction, we clarified which potential issues may hinder the success of online trade due to the asymmetry of information in possession of the parts involved in digital transactions. We distinguished Adverse Selection and Moral Hazard issues and we pointed out the role of feedback by previous users to reduce these information asymmetries. In this part we describe the types of feedback that digital platforms usually ask to users and report on their webpages. In particular, we focus on the nature of information that is usually displayed and the identity of the reviewers. At the same time, we sketch some of the main drawbacks associated with the online reviewing process such as review manipulation and reviewers' selection. Finally, we briefly illustrate the impact of reviews on users' online performance in terms of the volume of trade and prices.

eBay introduced its innovative review system in the year of its launch, 1995; with few modifications across the years, the same mechanism is still in use today. Later on, almost all digital marketplaces were inspired by the eBay feedback system and they implemented similar mechanisms with some adjustments due to the different contexts.

In general terms, review systems allow users to rate previous transactions with other parties with at least one numerical rating and one textual comment. The numerical rating can vary: in eBay, users can give a grade of $+1$, 0, or -1, while many other platforms use wider ranges (the five-star range seems to be the dominant choice across digital platforms). The text of the comments is usually restricted to few lines. Apart from the overall rating, users are commonly asked to review specific characteristics of the transactions with separate ratings: for instance, guests in Airbnb can separately review the location and the cleanliness of the hosts' dwellings; the dwellings' furniture; the accuracy of the webpage description; the hosts' communication skills and the check-in moment. All these ratings are then aggregated on the users' webpages with total and moving averages in order to facilitate the understanding of such a massive amount of information.

In almost all digital platforms, only subscribed users who had a reported transaction can review the other party. Many platforms use a bilateral reviewing process (eBay, Airbnb, BlaBlaCar) where the two parties review each other; while few marketplaces allow only one part to review the other: this is the case of Amazon where buyers can rate the sellers; but not vice versa.

TripAdvisor, Yelp and other interactive travel forums are noticeable examples of platforms that allow all website visitors to post reviews. Maizlyn, Dover and Chevalier (2014) show that the open structure of such review platforms facilitates reviews manipulation by third parties (such as competitors) and may lead to biased and incorrect representations of the quality of the services.

An additional source of reviews bias is associated with the users' fear of retaliation in some bilateral review systems: Klein, Lambertz, and Stahl (2016) and Fradkin, Grewal, and Holtz (2016) study these types of issues in eBay and in Airbnb, respectively. In both cases, authors argue that one party has incentives not to post negative reviews because of the risk to receive negative comments by the other party as a retaliatory behavior. These two studies show that, in absence of this risk, reviewers become less biased and report more often negative experiences.

The latter remarks about reviews accuracy give us the opportunity to recall three other main weaknesses of review systems:

- First, reviewing is almost always not mandatory and it greatly depends on the willingness to provide useful information to other users in the same community. Accordingly, only a part of the total number of users reviews and they may not be representative of the average users' tastes.

- Second, buyers experience may change over time because of the sellers' actions. Accordingly, past reviews may not be informative of the current level of the service quality.

- Finally, given the relatively low costs of creating accounts in digital platforms, users can delete their reviews' history after receiving bad comments; and start again with clean reputation.

The case of restaurant reviews illustrates all these points. In fact, skeptical readers of the online comments in TripAdvisor usually argue that those who review have very different tastes compared to their much more sophisticated palates; furthermore, the mood of restaurants staff changes from day to day and old reviews cannot capture this; finally, a perfect, but short reputation is suspicious and indicative of a recent cleaning of the online profile.[4]

The problem of reviewers' self-selection is difficult to eliminate or reduce with a modification of the reviewing process since it relates to the inner element of voluntary feedback mechanisms. Moreover, the potential bias related to the self-selection of users who decide to review may explain the great dominance of positive reviews in all digital platforms. Since reviewing is costly, only users who face extremely positive or negative experiences may decide to review. Alternatively, reviewers are self-selected among those who found a discrepancy between what they read in past feedback and the results of their own transactions: Dellarocas and Wood (2008) study these and other explanations for potential bias in eBay reviews. They conclude that eBay buyers who decide not to review have worse experiences. In line with this result, Nosko and Tadelis (2015) show that the ratio between positive reviews to the total amount of transactions is a more informative measure of the actual performance of eBay sellers. On top of this, social reciprocity may be an additional source for the positive bias of reviews in platforms where parties physically meet and the stakes of the services are higher, as Zervas, Proserpio, and Byers (2015) and Fradkin, Grewal and Holtz (2016) claim in the Airbnb case.

Despite their drawbacks, reviews do have an impact over users performance. In fact, in the last two decades several authors have investigated whether the reputation created by feedback systems matters and whether reviews have a significant bite in determining users' actions. Their findings differ depending on the platform and the type of empirical analysis. However, the most important studies agree in recognizing the following result.

[4] In spite of all these criticisms, Chua and Banerjee (2013) showed that TripAdvisor reviews are indeed largely reliable.

Finding 1. In several online platforms, the improvement of the users' reputation has a significant positive effect over users' number of transactions.

Here we will list relevant contributions on this topic focusing on robust results observed across several platforms using different methodologies. Cabral (2012) and Tadelis (2016) give excellent and comprehensive reviews of the most recent empirical literature on this topic.

The impact of online feedback over users' performance has been mainly documented on consumer-to-consumer (C2C) retail and e-commerce platforms such as eBay, Taobao and Amazon. A robust result across marketplaces regards the positive and significant effect of reviews on the volume of trade for sellers; instead, there is no complete consensus on the effect over prices.

The vast majority of studies focuses on C2C retail platforms where mostly non-professional sellers and buyers exchange goods: among them, eBay is the most studied case. Many scholars analyze how the buyers' reviews affect the outcome of future auctions for the sellers' objects. Dellarocas (2003) provides a complete summary of the first attempts to measure the effect of previous reviews on prices and probabilities of sale using cross-section regressions of sale prices on feedback. This approach has been discarded in most recent works starting with the article by Resnick *et al.* (2006): they use a field experiment and show that the results of previous cross-section analyses (a significant effect of reputation over sellers' performance) might be affected by the presence of omitted variables such as sellers' writing abilities. The authors randomly assign identical items (collector's postcards) to sellers with different reputations and they observe significantly higher winning bids for established sellers' accounts. Still, significant higher bids are also associated with those sellers who do not commit orthographic typos in the items' description.

To correct this bias, panel data analysis has replaced cross-section regressions: the article by Cabral and Hortaçsu (2010) is the most cited among those that apply panel data techniques. The authors construct a panel using feedback histories of several eBay sellers and focus on the impact of negative reviews over the weekly sales growth rates. They register a significant impact of the first negative review over the sales rate. From the movements of the sales rate and the amount of negative reviews they estimate the evolution of the sellers' behavior over time. Many other articles use panel data techniques to remove the confounding factors as the writing abilities evidenced by Resnick *et al.* (2006). Fan, Ju, and Xiao (2016) find returns to reputation in the Chinese platform Taobao. While established sellers result to have reputation

premia in terms of prices and volumes, new sellers with higher reputation tend to decrease prices to boost the sales rates further.

Anderson and Magruder (2012) and Luca (2011) use a different approach to evaluate the impact of restaurant reviews on the platform Yelp. Evaluating the impact of feedback over revenues presents one further issue in this context: restaurants with good reviews perform better than others because they are actually better. In this sense, observing a positive relationship between feedback and performance is not conclusive of the impact of feedback over performance. These types of problems are commonly referred to as reverse causality issues.[5] In both articles, the authors solve this issue implementing a regression discontinuity design: in the platform they study, users' ratings are aggregated and displayed on top of the restaurants' webpages as averages. These averages are rounded off to the nearest half-star (the rating range goes from one to five stars). In this sense, restaurants with very similar average ratings may have displayed a sensibly different number of stars on their webpages. For instance, a restaurant with an average rating of 4.2 appears to have four stars on its webpage; while a restaurant with 4.3 appears to have four stars and a half. Taking advantage of it, Anderson and Magruder (2012) and Luca (2011) compare restaurants with very similar underlying average ratings but with different displayed ratings and they estimate the effect of crossing the 0.5-stars on reservation availability (Anderson and Magruder, 2012) and revenues (Luca, 2011). In both cases reviews have a significant and positive impact.

The economic literature mainly focuses on numerical ratings; yet, textual comments constitute an important part of review systems since users may report essential pieces of information in the texts they write. Numerical ratings are bounded on a restricted range of values. Moreover, given the tendency of users to report positive reviews, the ratings' variance is often extremely small. By contrast, textual comments include a richer set of information and, if appropriately analyzed, they express a wider spectrum in users' experiences. Moreover, Filippas *et al.* (2017) show that textual comments in an online labor marketplace are less affected by review inflation, that is the tendency of users to lower their standards and give better feedback over time.

Finding 2. The significant impact of online reputation over users' performance is not restricted to numerical ratings, but it also regards textual comments.

[5] Reverse causality has a particular relevance in those contexts where reviews are not the unique source of information regarding the service and the evolution of ratings does not represent the only history of the transactions available. This is the case of travel forum websites such as Yelp, where users can find information about restaurants or other activities through many channels. In this sense, digital platforms such as eBay, or Airbnb are less affected by this issue.

This finding is supported by recent articles that explore textual comments with content or sentiment analyses: Ghose, Ipeirotis, and Sundararajan (2007) measure the strength and the polarity of comments in the Amazon review system and they study the economic impact of textual feedback over the performance of users. They observe that written reviews affect product sales and they measure how the comments' content determines the impact on users' performance in terms of sales. They find that reviews' characteristics such as subjectivity, readability and linguistic correctness influence sales and perceived usefulness of comments.

Archak, Ghose, and Ipeirotis (2011) identify different features of items sold in Amazon using a sentiment analysis of textual comments. Doing so, they are able to select the product features that consumers value the most and to analyze the reviews' impact over different product features. Their results show that textual reviews have an impact over prices and volumes of trade.

The empirical facts proposed by these papers show that reviews are important and buyers and sellers care about online reputation. In the next part we go beyond reviews' impact; and, in particular, we investigate how review systems are able to discipline the main issues related to Adverse Selection and Moral Hazard.

III. ADVERSE SELECTION: REVIEWS AS A SIGNAL FOR QUALITY

Asymmetric information between sellers and buyers is a feature that online exchanges share with many traditional markets. Accordingly, problems related to quality uncertainty are not new and many economists studied them years before the rise of digital trade. Akerlof (1970) introduces the concept of Adverse Selection and shows how buyers uncertainty regarding the quality of the objects sold in a market may lead to an (adverse) selection of the sellers who are willing to stay on the market and exchange. He studies cases in which buyers cannot apply any tools to objectively evaluate the quality of the goods on sale and shows as an example the market for used cars. In his article, buyers can only use prices to infer cars' quality and no mechanic tests are available. This total absence of methods to reduce the uncertainty on the buyers' side may be too restrictive since certifications and warranties are often present in reality to evaluate the quality of products. In fact, many works show that these tools can help to reduce the asymmetry of information.[6] In online markets, reviews play a role similar to certifications in that they provide additional information about the quality of the items listed on the platforms;[7] and reviews can be

[6] Dranove and Jin (2010) provide a complete review about the efficacy of these tools.

[7] Elfenbein, Fisman, and McManus (2015) show that eBay feedback serves as a substitute for eBay's own quality certification.

considered as a "signaling device" to learn the quality of the object, as pointed out by Dellarocas (2006).

Some theoretical works investigate how the observation of outcomes of past transactions can foster buyers' learning about sellers' quality.[8] Since the outcomes of transactions may be a noisy measure of the actual seller quality, potential buyers need many observations to fairly infer the quality: with an infinite amount of observations, buyers learn perfectly. Yet, the flow of observations may stop before inducing a sufficient learning of the true quality of sellers, who may exit the market irrespectively of their quality. This may be the case of online high-quality sellers who were unlucky in the very first transactions and received bad reviews. Because of this effect no buyer is willing to purchase their items, keeping their (bad) reputation not updated.

Bar-Isaac (2003) shows the important role of the seller's belief about his own quality. If a seller knows his quality, then learning failures are less common since good-quality sellers may decide to stay and decrease the price they charge in case of a temporary bad reputation. The future profits obtained after the true (good) reputation is restored can compensate the losses made in the first periods with bad reputation. Conversely, if a seller does not know his quality, buyers' reviews shape the seller's beliefs regarding his own quality: a few bad reviews may convince the seller to be of low-quality and induce him to exit since he does not expect better reviews in the future.

Finding 3. Online reviews foster buyers' learning of sellers' quality reducing Adverse Selection. Still, learning may stop irrespectively of the true sellers' quality if users' reviews are a noisy measure of the true quality.

From an empirical point of view, cases of learning failures are difficult to observe since quality is sellers' private information. Still, the learning patterns evidenced before can be observed in the studies of several scholars who investigated the effect of reviews on sales in the movie industry: in this setting, quality is fixed over time and online reviews are a noisy measure of quality since they are affected by users' tastes. We will list here a few papers that study how the word-of-mouth expressed by online feedback influences the movies' box office performance. Dellarocas, Zhang, and Awad (2007) build an econometric model to forecast the dynamics of movies' box office revenues over time. Their model includes, as predictors, pre-release marketing, professional critic reviews and the number of theaters where the movies were shown. They observe that users' reviews published on several review aggregation websites (Yahoo!Movies, BoxOfficeMojo and the Hollywood Reporter) improve the forecasting ability of

[8] Bar-Isaac *et al.* (2008) provide an excellent summary of such models.

the model and show that online word-of-mouth has a significant bite over the movies' sales. Duan, Gu, and Whinston (2008) study the dynamic relationship between sale volumes and reviews approaching the reverse causality issue introduced in the previous section: with a dynamic simultaneous equation system they show that the volume of online reviews improves the box office performance of movies. Furthermore, movies' box office revenues also increase word-of-mouth volume, creating a reinforcing dynamics between sales and reviews in line with the process of learning highlighted in the theoretical works described above.

Since the issues regarding learning of sellers' quality is of great importance for online trade, some digital marketplaces implemented particular mechanisms to induce the correct learning of the sellers' quality; and thus to diminish the market inefficiencies due to Adverse Selection. One of the most studied tools to signal quality in review systems is the possibility that sellers provide incentives for buyers to leave feedback. The Chinese C2C platforms Alibaba and Taobao launched in the recent years a feedback reward mechanism called "Rebate-for-Feedback" (RFF) for online sellers. When sellers choose this option, they set a rebate amount for any item they sold to buyers conditional on buyers leaving highly informative feedback. The informativeness of the feedback is computed with a machine-learning technique programmed by the platforms. High-quality sellers who know their quality and have recently entered the platforms have incentives to use RFF for two main reasons: first, buyers have incentives to leave a descriptive feedback of their (high) quality and the learning process will speed up. Second, buyers know whether sellers opted for the RFF feature and they may consider this as a signal for quality since the sellers want to be reviewed. Li (2010) shows with a theoretical model that this type of mechanism can reduce Adverse Selection as well as the bias of reviews since a wider range of users will review. Even though both high-quality and low-quality sellers choose this option in equilibrium, buyers prefer sellers who choose it and their true types are revealed through feedback. Li and Xiao (2014) test the predictions of this model in a lab experiment and they find a consistent evidence; Cabral and Li (2015) study a similar mechanism with a monetary reward of feedback using a series of field experiments in eBay. They observe buyers leaving more and better feedback for those sellers who give monetary rewards. Finally, Li, Tadelis, and Zhou (2016) study the RFF mechanism using Taobao data and show that sellers who choose this option have higher sales and better feedback with respect to those who do not choose it. This suggest that RFF can be considered as a signal for quality that buyers understand; and a useful tool to fight Adverse Selection.

Finding 4. The power of review systems to signal sellers' quality can be improved with incentives for buyers' to report their feedback so as to reduce learning failures and improve the informativeness of reviews.

Coming back to the cases where no signaling devices are present apart from reviews, some recent articles study the entry and exit dynamics of sellers when reputation determines the beliefs over their quality and the prices they can charge. Atkeson, Hellwig, and Ordoñez (2014) focus their analysis on the role of entry taxes over these dynamics. The authors assume that, before entering the market, sellers can invest in their own quality, that remains fixed after entry. Entry taxes create incentives for sellers to invest: accordingly, sellers' entry reputation increases and the informativeness of reputation will be reinforced. Vial and Zurita (2017) add to this framework the possibility for sellers to change names over time and start with a new (clean) reputation. Studying name changing strategies is extremely important since this behavior can harsh the entire functioning of feedback mechanisms. In the model by Vial and Zurita (2017), the starting reputation of new entrants (those with no reviews at all) plays a key role since sellers with lower reputation than entrants decide to change name. Their model predicts well the major empirical findings of the literature with "younger" sellers being more likely to exit (that is, starting with clean records) and the probability to exit increasing as reputation worsens.

The empirical literature about the reviews' impact on sellers' performance is in line with the idea of users learning the quality through past feedback. In this sense, the positive impact of ratings over sales rates is due to the change in buyers' beliefs regarding the sellers' quality; prices accommodate changes in reputation since buyers expect different qualities from different reputation levels. Studies about the relationship between ratings and prices have to take into account the multiple channels that link these two variables; and how movement in prices may be used by sellers to induce further learning of their quality.

Jolivet, Jullien, and Postel-Vinay (2016) report a significant effect of reputation over prices in the e-commerce platform PrimeMinister and explicitly consider the dynamic relationship between prices and ratings: better reputation leads to higher prices; still, high prices may increase buyers' expectations and potential dissatisfaction.

Fan, Ju, and Xiao (2016) analyze how sellers manage their reputation through the life cycle in the Chinese platform Taobao. They distinguish between new and experienced sellers and show that the effects of reputation for these two classes of users are different: new sellers do not increase prices after receiving the first positive reviews. But, they keep them low to further boost their volumes of trade. After many reviewed transactions, new sellers become experienced sellers, with a stronger reputation and the possibility to exploit the reputation to increase prices.

However, prices are not the only variable that determines buyer's value of a transaction. In almost all digital platforms, sellers can affect the quality of the services over time through effort. In the next part, we focus on Moral Hazard issues: first, we illustrate how reviews can be used as an on-going monitoring device of the behavior of users in digital marketplaces; moreover, we discuss theoretical and empirical works related to these contexts.

IV. MORAL HAZARD: AVOIDING MISBEHAVIOR WITH REVIEWS

Together with Adverse Selection, Moral Hazard issues are common features of traditional and digital marketplaces. In several cases agents have no incentives to perform well in one-shot interactions; still, if agents interact in several periods, incentives against misbehavior can arise. With repeated interactions, agents' misconduct today may lead to punishment tomorrow; while cooperating today may lead to future rewards. Game theory studies these cases. In particular, one of the most remarkable results of this field (called the Folk Theorem) shows that, with a sufficiently high discount factor, any outcome, also very beneficial for all parties, can be sustained in equilibrium.[9]

This conclusion can be applied to a basic game where one seller and one buyer repeatedly trade with the following timing: first, the buyer can send or not money to the seller in exchange of a good; next, when the seller receives the monetary transfer, he decides whether to send the good or not. With trade occurring only once, the seller never sends the good after receiving the money and he keeps the object for his personal use. Accordingly, the buyer never sends the money since he cannot trust the seller: agents do not trade.

Still, when this game is repeated over time, the buyer may apply a trigger strategy: he sends money each period until the seller stops sending the object. When it happens, he stops sending money. With this strategy, the seller decides to send the object if future profits from trades exceed the value of keeping the objects for his personal use today and in the future. In this case, buyer and seller will trust each other and they will trade in each period. However, if trade can take place only during a finite number of periods and agents know when exchanges end, trust between buyer and seller cannot be built: in the very last period, agents are back in the same situation of the static game and they will

[9] The entry in *The New Palgrave Dictionary of Economics* by Kandori (2008) contains an instructive review over the studies about repeated interactions with a game theoretic approach.

not trade. Applying the same argument, trust in any period before the last one cannot be sustained.

Yet, users in digital markets rarely interact multiple times and it is reasonable to assume that they do not know each other before trade. Still, thanks to review systems online buyers can observe the outcomes of the previous transactions of a seller and notice whether he is trustworthy or not: in the previous example, if a seller always shipped the object or not. Therefore, thanks to the presence of past reviews, it is possible to build trust with trigger strategies played by all the sequence of buyers who have transactions over time with the seller: buyers start sending money and write positive reviews after receiving the good; then, if once the seller does not send the object, the buyer will write a negative review and all the next buyers will know about the seller misbehavior; hence, they will stop sending money.

However, as we pointed out before, if the seller knows that he is going to exit the market for sure at a certain date, then his incentives to behave properly in the last transaction decay and misconducts can arise.

Finding 5. Through past reviews, buyers can monitor seller's past behavior. Sellers have incentives to behave correctly since, in case of misconduct, buyers will punish them with negative reviews.

Empirical studies find that punishment and rewards strategies are at play from the buyers' side in several online platforms. Still, in reality, online buyers do not implement pure trigger strategies that would lead to a complete cease of the sellers' activities after a negative review. Additionally, reviews are not perfectly informative about the quality of the transactions because of the multiple sources of review bias expressed in the previous part.

In eBay, Cabral and Hortaçsu (2010) find that sellers' sales significantly drop (from 5% to -8%) after the first negative review.

Moreover, seller behavior changes depending on his reputation: in the same article, Cabral and Hortaçsu (2010) report that after the first negative rating, further negative feedback follows 25% more frequently; still, with a lower impact on the sellers' performance. With high reputation, the incentives to behave well are also high; conversely, if the level of reputation goes down because of a negative review, then sellers are less motivated to perform well. Cabral (2015) proposes a theoretical model for this type of behavior that can explain the persistence of high performance of online traders.

Finding 6. When sellers plan to exit the platform, the incentives for good conduct provided by review systems are weak: the majority of negative reviews occurs close to the end of sellers' life-cycle.

If a seller knows that he is going to exit soon, then future profits from good behavior reduce and cases of misconduct are more likely. This theoretical finding is in line with the empirical evidence. Cabral and Hortaçsu (2010) show that the lower the sellers' reputation, the higher their exit probability; and sellers receive more negative reviews before exiting than in their lifetime average. Still, the relationship between exit and negative reviews is also in line with another story: the performance of a seller may be reviewed badly for external reasons to the effort he puts (in the previous example, a seller can ship the objects, but buyers never receive them because of postal disservice). Due to this, seller's reputation decreases and he prefers to exit rather than exerting effort to recover a good reputation.

Following the theoretical predictions and the empirical findings, we may conclude that sellers' and buyers' strategies evolve over time as information about the transactions slowly accumulates on their webpages. Newcomers on the platforms have more incentives to behave well and build a positive reputation. Whereas later they enjoy high reputation and profitable exchanges. Finally, closer to the exit, sellers' incentives to misbehave are higher and they will end up their life-cycle on the platform with a higher rate of negative reviews.

V. ADVERSE SELECTION, MORAL HAZARD AND REVIEW SYSTEMS: A GENERAL OVERVIEW

In the previous sections of this chapter we described review systems and we showed evidence of their impact on online buyers and sellers. Later, we analyzed separately Adverse Selection and Moral Hazard introducing a theoretical framework and the empirical findings corroborating the theories. In this way, the signaling and sanctioning functions of online feedback have been enlightened with several examples. At the same time, these two issues are closely related and different theories may explain the same empirical facts. An example of these similarities was given by the two theories that motivate the relationship between sellers' exit decisions and a drop in their reputation in the last periods of their stay on the platform. The bad reputation of sellers may be related with the inner qualities of sellers' services. This explanation is more in line with Adverse Selection and the learning process described in the third part of this chapter. At the same time, exit decisions by sellers may correlate with

bad reputation because of a drop in the sellers' effort; and Moral Hazard issues are in place.

In the remaining part of this chapter, we analyze Adverse Selection and Moral Hazard together; first, we present two alternative theoretical contributions dealing with Adverse Selection and Moral Hazard when users interact over time and they can build a reputation from the reviews of previous transactions. Afterwards, we discuss the presence of these two issues in several platforms pointing out how the interpretation of Adverse Selection and Moral Hazard can vary across digital contexts. Finally, we focus on some recent empirical works, in line with the models presented, that study how changes in review systems design can reduce Adverse Selection and discipline Moral Hazard.

The discussion of the theoretical models follow the excellent review by Bar-Isaac *et al.* (2008) where the authors analyze these and other types of models regarding seller reputation.

In the previous models of learning and repeated interactions, the buyers' uncertainty regards either the fixed quality of the seller; or, the seller's decisions in each trade event. Other models extend the previous frameworks and discuss cases where sellers' quality and decisions are unknown to buyers at the same time. We start analyzing the "signal jamming" model presented by Holmström (1999). In this model, a manager works in each period for a different company and his performance with the companies can result either in a success, or in a failure. The probability to be successful in each period depends on the sum of two elements: manager's innate ability and effort. The innate ability is unknown to the companies and to the manager. Still, the manager can choose the effort to put in each period and everybody observes the history of manager's successes or failures in previous transactions. Moreover, companies pay a wage to the manager in line with the expected probability of success that they infer from the history. The manager's objective is to achieve the highest lifetime wages minimizing the effort.

Holmström (1999) shows that, in equilibrium, the manager chooses high effort in the first transactions to influence the companies learning process, and the associated wage process. Still, the effort diminishes over time since, in the long run, companies perfectly infer the manager's ability and they pay him a wage based on his ability. Accordingly, the model explains the career concerns of agents who exert high effort at the beginning of their working life, lowering their care in performing well when the reputation is built.

This framework perfectly fits the case of online trade with one seller trading each period with different buyers who observe the outcomes of previous

transactions thanks to the reviews. Moreover, the theoretical findings of a decreasing effort over time are in line with the empirical facts about the life cycle of eBay sellers reported by Cabral and Hortaçsu (2010).

In the previous model, quality and effort sum together to form the expected productivity of the manager. However, we may interpret the concept of quality as the capacity of sellers to perform well exerting the necessary effort for the transactions. In this sense, high-quality sellers are those that do not act strategically and always ship the objects to buyers. Differently, low-quality sellers can change their shipping decisions over time, with potential misconducts. Kreps *et al.* (1982), Kreps and Wilson (1982) and Milgrom and Roberts (1982) introduce different types of sellers inside the framework of repeated games. In their analysis, they consider two types of sellers: a commitment type and a strategic type. Commitment types are always playing the action to which a long-run player would like to commit: that is, exerting high effort in all the transactions. Conversely, strategic types are not constrained in their decisions and they can choose in each period whether to put effort, or not.

To explain the economic rationale of these models, we refer back to the basic game between a seller and multiple buyers illustrated in the previous part. Now the seller can be either a commitment type and he will always send the object; or, a strategic type and he will choose to ship the object or not in each period. Buyers do not know the type of the seller, but they are aware that commitment and strategic sellers are both present on the platform. In this sense, the history of previous transactions has a double function for buyers: past reviews help to monitor the on-going behavior of the seller as in the previous case without multiple types of sellers; furthermore, they may signal the type of seller. If reviews are perfectly representative of the quality of transactions, then commitment types always face positive reviews and buyers can infer the strategic nature of sellers with only one negative reviews. Because of this, strategic players have incentives to always ship the object to buyers acquiring the reputation of a commitment type.

Allowing feedback to be a noisy measure of the sellers performance, such a direct inference is no longer valid since also commitment types may be "unlucky" and get negative reviews. Cripps, Mailath, and Samuelson (2004) show that in this case, strategic types do not always imitate commitment: after having established a good reputation with many positive reviews, strategic types may not send the object in some transactions blaming external factors involved in the shipping. In the long run, types will be learned and reputation concerns disappear.[10]

[10] Situations with other seller types may originate different results regarding the impact of reputation. Bar-Isaac *et al.* (2008) and Cabral and Hortaçsu (2010) extensively review all these models.

These two classes of models study how reputation affects sellers' behavior when buyers' uncertainty regards the fixed quality and the decisions of sellers over time. Some predictions of the evolution of sellers' actions are common: reputation effects are strong in the initial phase of sellers' life cycle; and decreasing over the number of transactions. Yet, some relevant differences are present regarding how the two types of uncertainty are related. In Holmström (1999), the innate ability and the effort of the manager play the same role in determining the probability of success and the manager quality does not affect directly the effort decisions; we have to recall that the manager is not aware of his innate ability and he learns it with the companies from the history of performance. Differently, the literature about seller types in repeated games defines the quality of a seller as his capacity to act in a non-strategic way. This distinction is not only important from a theoretical point of view, but it interests the nature of the services enabled by different digital platforms.

In sharing-economy platforms such as Airbnb, BlaBlaCar, or TaskRabbit, the quality of services provided is composed by a part that is fixed over time, and by the time-varying attention and care of users. For instance, the quality of a stay in a house listed on Airbnb depends at the same time on the dwelling's quality (that may be fixed, as the dwelling's location) and on the hosts' attention in cleaning, communicating and receiving the guests. Accordingly, the model by Holmström (1999) has a better fit for these types of platforms as suggested by the empirical findings presented by Rossi (2018) regarding Airbnb. In his work, a sentiment analysis of guests' comments is used to disentangle two dimensions of the quality of hosts' service: one dimension regards how guests evaluate the fixed component of the service due to the dwelling's quality. The other dimension relates to the guests' perception of the hosts' effort. Both measures include an amount of "noise" due to the tastes and perceptions of guests. To remove the guest idiosyncratic component, Rossi (2018) uses a control function approach that establishes a relationship between the guests' tastes about the dwelling's quality and the hosts' effort. Having removed the idiosyncratic guests' perceptions, an estimate of the dynamics of the effort exerted by Airbnb hosts over time is obtained. In line with the model by Holmström (1999), Airbnb hosts exert a higher effort in the first transactions to attract guests; while they shirk in the transactions before exit since the reputational incentives are low.

The case of C2C and e-commerce marketplaces is different: here it is hard to distinguish between fixed and varying aspects of the exchange quality. A high-quality seller is the one who describes properly the state of his goods, and respects the delivery deadlines. Even though sellers may change their policies over time, we may consider these behavioral features as fixed over time for some sellers. In this fashion, models with different types of sellers that trade

repeatedly with buyers are more often used to explain the empirical findings regarding the sellers' behavior in these platforms.

Finding 7. Irrespectively of the type of model, when reviews are signals for quality and sanctioning devices, two results emerge: 1) users learn the true value of sellers' quality after a sufficient number of reviews; 2) reputation incentives for good behavior are stronger at the beginning of the life-cycle and weaker close to exit.

We conclude this part discussing some empirical papers that exploit variations of review systems design to observe how these changes impact on Adverse Selection and Moral Hazard.

Klein, Lambertz and Konrad (2016) and Hui, Saeedi, and Sundaresan (2017) take advantage of a variation in the eBay review system implemented in 2008 to remove the potential bias of feedback due to the buyers' fear of retaliation. In both studies the authors observe that the variation led to a significant reduction of the inefficiencies due to asymmetric information; still, Klein Lambertz and Konrad (2016) claim that it induced a disciplining effect on Moral Hazard; instead, Hui, Saeedi and Sundaresan (2017) attribute the improvement to a reduction in Adverse Selection. Here we compare their methodologies and their results.

It has been shown that many eBay users, before starting selling objects, decide to build a reputation as buyers. This behavior was firstly noticed by Cabral and Hortaçsu (2010) and several other articles confirm the same empirical fact. Accordingly, eBay buyers care about their reputation in that they will use it later when they start their career as sellers. Before 2008, eBay sellers, in case of buyers' negative reviews, were used to retaliate with negative reviews: evidence of this is provided by Hui, Saeedi and Sundaresan (2017), who report that sellers responded with negative feedback after receiving negative feedback from buyers in the 37% of the cases. This retaliatory behavior, together with the interest of buyers in keeping a good reputation, created a positive bias over reviews with buyers under-reporting sellers misconduct. To eliminate this bias, eBay modified in May 2008 its feedback process allowing sellers to rate buyers only with positive reviews (or no feedback).

Klein, Lambertz and Konrad (2016) evaluate the impact of this change in the eBay review process. They compare the levels before and after May 2008 of the Detailed Seller Ratings (DSRs), the anonymous feedback that buyers can

report after each transaction; and, of the sellers' exit rate. They find that the change led to a significant improvement in DSRs. Since this type of rating has always been anonymous, they infer that it has never been biased by the fear of retaliation and buyers' satisfaction improved after the change. Differently, the exit rate of sellers is not affected.

In this sense, their results suggest that the feedback variation disciplines Moral Hazard; that is, sellers behave better after May 2008. However, it does not lead to a reduction of Adverse Selection since sellers exit rate does not increase.

In contrast with this study, the empirical findings by Hui, Saeedi and Sundaresan (2017) are more in line with a reduction in Adverse Selection. To measure the movements in sellers' quality before and after the change they study several parameters: negative feedback (not anonymous); DSRs (anonymous); and the number of buyers' disputes. In addition, they consider the sellers' size, that is, the number of items sold in a given month; and the sellers' exit rate. They measure the change in buyers' satisfaction due to changes in sellers' behavior and changes in the sellers' size; and they interpret the former as a reduction in Moral Hazard and the latter as a reduction in Adverse Selection. Doing so, they estimate that the reduction of Adverse Selection accounts for the 68% of the buyers' satisfaction improvement. While the discipline of Moral Hazard accounts for the remaining 32%.

The opposite conclusions by Klein, Lambertz and Konrad (2016) and Hui *et al.* (2017) are probably due to the different nature of the datasets used by the authors. In particular, Hui, Saeedi and Sundaresan (2017) used eBay proprietary data, while Klein, Lambertz and Konrad (2016) scraped data from the eBay website. As suggested by Hui, Saeedi and Sundaresan (2017), using scraped datasets may bias the results in that the eBay sellers studied by Klein, Lambertz and Konrad (2016) are seasoned sellers who stay active on eBay for more than a year and whose probability of exiting the platform is much lower than average.

Fradkin, Grewal and Holtz (2016) analyze the effect of a similar variation in the Airbnb review system. In this platform, having a bilateral feedback system is necessary because of the significant uncertainty regarding the profiles of guests and hosts. In this sense, Airbnb has not modified the two-sided design of its review system (as eBay did in May 2008); but, to avoid retaliation, hosts and guests reviews are posted simultaneously on users' webpages after the change in May 2014. Fradkin, Grewal and Holtz (2016) study the outcomes of several experiments that led to the adoption of such a policy by Airbnb using proprietary data. They show that the simultaneous reveal experiments increase

review rates leading to a more precise learning of users' quality and improving market efficiency.

We conclude this review of empirical works with the article by Hui, Saeedi and Sundaresan (2016) where they discuss jointly the roles of reputation and regulation in reducing asymmetric information. In this paper, the authors focus on two programs by eBay: the Top Rated Seller (TRS) program, implemented in October 2009; and the Buyer Protection (BP) program, active from October 2010. The TRS identifies the most reliable sellers considering their past performance and sales volume. Top Rated sellers are signaled with a badge shown on top of the eBay webpage. Differently, the Buyer Protection program aims at guaranteeing purchases from all sellers. Thanks to the BP program sellers have to refund buyers if the items are not received; or if the items differ from the ones described online.

First, Hui, Saeedi and Sundaresan (2016) establish that the TRS badge has a positive signaling value for sellers since the average sales price for sellers that are badged raises by 3%. Moreover, badged sellers perform better than those who are not badged.

Later, they study the regulatory effect of the BP program. They show that negative feedback ratings decrease by 23% after the introduction of the program. Thus, they conclude that the regulation provided by the BP program had a significant impact on Moral Hazard. Moreover, the quality of eBay sellers increases with a reduction of Adverse Selection: the exit rate for low quality sellers increases as well as the share of Top Rated sellers.

The brief overview on recent articles captures, at least partially, the state of the art regarding how the fine-tuning of review systems affects the asymmetry of information due to Adverse Selection and Moral Hazard. The following finding summarizes the main results.

Finding 8. More accurate reports on seller behavior (with lower fear of retaliation from the buyers' side) reduces asymmetry of information in two ways: 1) It mitigates Adverse Selection since low-quality sellers exit or their sales' volume shrinks; 2) It disciplines Moral Hazard since buyers are free to punish sellers in case of misconduct. Moreover, digital platforms may jointly rely on reputation (using reviews) and regulation (using guarantees and certifications) to improve the quality of the services provided and to reduce the asymmetry of information.

VI. CONCLUSION

In this last part we conclude with a recap of the most important points analyzed; and with a list of further directions of research regarding these issues.

This chapter aims at clarifying the role of review systems in reducing asymmetric information in digital platforms. When the phenomenon of e-commerce and digital trade started, experts were alarmed by some features that could severely hinder the existence and the efficiency of these markets. In the introduction, we grouped all these criticisms in two parts: online buyers do not perfectly know the quality of sellers and this uncertainty may adversely select the sellers. At the same time, sellers exert effort once buyers have paid for the transaction; hence, Moral Hazard issues may be at play.

Next, we described the common design of review systems in digital plaforms and we illustrated possible weaknesses of the mechanisms currently adopted in online marketplaces. Despite these shortcomings, online reputation matters and online users care about reviews: this result is observed in several platforms and using different techniques.

After a brief review over the impact of feedback on users' performance, we discussed the theoretical mechanisms and empirical findings on how reviews of past transactions can reduce Adverse Selection and discipline Moral Hazard.

- First, we considered the role of reviews in signaling sellers' quality and circumstances in which buyers' learning process stops (Bar-Isaac, 2003). With this respect, we reviewed theoretical and empirical studies in favor of a mechanism implemented by two Chinese platforms: the Rebate-for-Feedback.

- Later, we focused on Moral Hazard describing the theoretical mechanisms to create incentives for sellers' good behavior when transactions are repeated.

- Finally, we described two theoretical models that consider Adverse Selection and Moral Hazard simultaneously. After discussing the applications of these models in different contexts, we listed some recent empirical works that identify the impact of reviews in reducing the asymmetries of information exploiting variations in the feedback design.

The literature about digital markets and reputation keeps growing rapidly. We suggest here some potential directions of future research in this field. Our

short list of possible avenues of research is not exhaustive and for the advanced readers we suggest the excellent works by Dellarocas (2003) and Cabral (2012).

Users' behavior in digital platforms presents many unanswered questions: why do users review? What do they review? Reviews are a public good and they provide positive externalities to the users' community. Still, reviewing has a cost and, from a pure economic point of view, users have no incentives to leave their feedback. Only recently Fradkin, Grewal and Holtz (2016) and Filippas *et al.* (2017) have opened the discussion over these issues; still, given the relevance of these questions, further research is necessary from a theoretical and empirical perspective.

A second promising line of research is related to the emergence of new types of platforms associated with the sharing economy:[11] these marketplaces connect people and favor exchanges with higher stakes relative to C2C or e-commerce websites. Accordingly, mechanisms to ensure services' quality such as review systems and regulations are particularly important for the success of these platforms. Still, only few works have studied these contexts, observing that reviews are important (Edelman, Luca, and Svirsky, 2017) and additional evidence is required to establish robust results. Moreover, both Adverse Selection and Moral Hazard issues are potentially present in many services that are offered on these marketplaces. Time-varying effort affects the quality of the exchanges as well as the characteristics of some facilities that are fixed over time. In this sense, sharing economy platforms such as Airbnb, BlaBlaCar, or TaskRabbit are an ideal setting to test the predictions of models of reputation where Adverse Selection and Moral Hazard are both present and to understand how fixed characteristics and effort are related. The work by Rossi (2018) investigates these issues in the Airbnb setting. The dynamics of the effort exerted by Airbnb hosts are only partialy influenced by the quality of their dwellings. Hosts tend to exert high effort at the beginning of the life-cycle and shirk close to the end independently of the house's quality. Still, hosts with low-quality dwellings stay for shorter periods on the platform with sharper changes in hosts' effort over the life-cycle.

Finally, there is no consensus about the characteristics of an "optimal" feedback mechanism that is free from the shortcomings previously listed. Which changes in review systems are needed to facilitate trust?

On the empirical side, the introduction and the positive impact of mechanisms such as the Rebate-for-Feedback and the Buyer Protection programs show how the proper design of review systems leads to a significant reduction

[11] Sundararajan (2016) provides an extensive overview on the economics of these platforms and the main issues related to the growth of the crowd-based capitalism.

of inefficiencies. In this sense, further works are important to understand what programs are more effective in different contexts.

From a theoretical point of view, Dellarocas (2005) pioneered the normative approach about the design of a reputation mechanism to discipline Moral Hazard. Along the same lines, Aperjis and Johari (2010) and Bolton, Greiner, and Ockenfels (2013) investigate the optimal pieces of information that platforms should show and aggregate to facilitate trust among users, signal the users' quality and create incentives for good behavior. However, until now no general consensus has been achieved in the theoretical literature regarding the selection of the most relevant information that review systems should provide in contexts with different degrees of Adverse Selection and Moral Hazard.

With this chapter, we give a systematic overview of the theoretical and empirical works related to the issues of asymmetries of information in digital contexts and the role of review systems. Recalling the anecdote of the broken laser pointer in the very first eBay transaction, the well-functioning of online operations was not obvious even for the founder of the first successful digital marketplace. Whereas now, digital platforms connect millions of users daily and the possibility to trade safely online is no more under question. For sure, one reason of the great success of online markets is the introduction of innovative review systems that helped to discipline users' behavior and signal their quality.

BIBLIOGRAPHY

Akerlof, G. A. (1970), "The market for "lemons": Quality uncertainty and the market mechanism," *The Quarterly Journal of Economics:* 488–500.

Anderson, M., and J. Magruder (2012), "Learning from the crowd: Regression discontinuity estimates of the effects of an online review database," *The Economic Journal,* 122(563): 957–989.

Aperjis, CH., and R. Johari (2010), *Designing reputation mechanisms for efficient trade.*

Archak, N.; Ghose, A., and P. G. Ipeirotis (2011), "Deriving the pricing power of product features by mining consumer reviews," *Management Science,* 57(8): 1485–1509.

Atkeson, A.; Hellwig, CH., and G. Ordoñez (2014), "Optimal regulation in the presence of reputation concerns," *The Quarterly Journal of Economics,* 130(1): 415–464.

Bar-Isaac, H. (2003), Reputation and survival: Learning in a dynamic signalling model, *The Review of Economic Studies,* 70(2):231–251.

Bar-Isaac, H. *et al.* (2008), "Seller reputation. Foundations and Trends®," *Microeconomics,* 4(4): 273–351.

Bolton, G.; Greiner, B., and A. Ockenfels (2013), "Engineering trust: reciprocity in the production of reputation information," *Management Science,* 59(2): 265–285.

Cabral, L. (2012), "Reputation on the internet," *The Oxford Handbook of the Digital Economy:* 343–354.

— (2015), "Living up to expectations: Corporate reputation and persistence of firm performance," *Strategy Science,* 1(1): 2–11.

Cabral, L., and A. Hortaçsu (2010), "The dynamics of seller reputation: Evidence from ebay*," *The Journal of Industrial Economics,* 58(1): 54–78.

Cabral, L., and L. Li (2015), "A dollar for your thoughts: Feedback-conditional rebates on ebay," *Management Science,* 61(9): 2052–2063.

Chua, A. Y. K., and S. Banerjee (2013), "Reliability of reviews on the internet: The case of tripadvisor," *Proceedings of the World Congress on Engineering and Computer Science,* volume 1.

Cohen, A. (2003), *The perfect store: Inside eBay,* Back Bay Books.

Cripps, M. W.; Mailath, G. J., and L. Samuelson (2004), "Imperfect monitoring and impermanent reputations," *Econometrica,* 72(2): 407–432.

Dellarocas, CH. (2003), "The digitization of word of mouth: Promise and challenges of online feedback mechanisms," *Management Science,* 49(10): 1407– 1424.

— (2005), "Reputation mechanism design in online trading environments with pure moral hazard," *Information Systems Research,* 16(2): 209–230.

— (2006), "Reputation mechanisms," *Handbook on Economics and Information Systems:* 629–660.

Dellarocas, CH., and CH. A. Wood (2008), "The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias," *Management Science,* 54(3): 460–476.

Dellarocas, CH.; Zhang, X. M., and N. F Awad (2007), "Exploring the value of online product reviews in forecasting sales: The case of motion pictures," *Journal of Interactive marketing,* 21(4): 23–45.

Dranove, D., and G. Z. Jin (2010), "Quality disclosure and certification: Theory and practice," *Journal of Economic Literature,* 48(4): 935–963.

Duan,W.; Gu, B., and A. B. Whinston (2008), "The dynamics of online word-of-mouth and product sales an empirical investigation of the movie industry," *Journal of retailing,* 84(2): 233–242.

Edelman, B.; Luca, M., and D. Svirsky (2017), "Racial discrimination in the sharing economy: Evidence from a field experiment," *American Economic Journal: Applied Economics,* 9(2): 1–22.

Elfenbein, D. W.; Fisman, R., and B. McManus (2015), "Market structure, reputation, and the value of quality certification," *American Economic Journal: Microeconomics,* 7(4): 83–108.

Fan, Y.; Ju, J., and M. Xiao (2016), "Reputation premium and reputation management: Evidence from the largest e-commerce platform in China," *International Journal of Industrial Organization,* 46: 63–76.

Filippas, A., *et al.* (2017), Reputation in the long-run, Technical report, CESifo Group Munich.

Fradkin, A.; Grewal, E., and D. Holtz (2016), The determinants of online review informativeness: Evidence from field experiments on airbnb. Technical report, *Working Paper.*

Ghose, A.; Ipeirotis, P., and A. Sundararajan (2007), Opinion mining using econometrics: A case study on reputation systems, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics:* 416–423.

Holmström, B. (1999), "Managerial incentive problems: A dynamic perspective," *The Review of Economic Studies,* 66(1): 169–182.

Hui, X.; Saaedi, M.; Shen, Z., and N. Sundaresan (2016), "Reputation and regulations: evidence from ebay," *Management Science,* 62(12): 3604–3616.

Hui, X.; Saeedi, M., and N. Sundaresan (2017), Adverse selection or moral hazard: An empirical study, *Working paper,* 1-46.

Jolivet, G.; Jullien, B., and F. Postel-Vinay (2016), "Reputation and prices on the e-market: Evidence from a major french platform," *International Journal of Industrial Organization,* 45: 59–75.

Kandori, M. (2008), Repeated games, *New Palgrave Dictionary of Economics, 2nd edition, Palgrave Macmillan.*

Klein, T. J.; Lambertz, CH., and O. S. Konrad (2016), "Market transparency, adverse selection, and moral hazard," *Journal of Political Economy,* 124(6): 1677–1713.

Kreps, D. M., and R. Wilson (1982), "Reputation and imperfect information," *Journal of Economic Theory,* 27(2): 253–279.

Kreps, D. M.; Milgrom, P.; Roberts, J., and R. Wilson (1982), "Rational cooperation in the finitely repeated prisoners' dilemma," *Journal of Economic theory,* 27(2): 245–252.

Li, L., and E. Xiao (2014), "Money talks: rebate mechanisms in reputation system design," *Management Science,* 60(8): 2054–2072.

Li, L. I. (2010), "Reputation, trust, and rebates: How online auction markets can improve their feedback mechanisms," *Journal of Economics & Management Strategy,* 19(2): 303–331.

Li, L. I.; Tadelis, S., and X. Zhou (2016), Buying reputation as a signal of quality: evidence from an online marketplace, Technical report, National Bureau of Economic Research.

Luca, M. (2011), Reviews, reputation, and revenue: The case of yelp. com. Com (September 16, 2011), *Harvard Business School NOM Unit Working Paper,* (12-016).

Manyika, J.; Lund, S.; Bughin, J.; Woetzel, J. R.; Stamenov, K., and D. Dhingra (2016), *Digital globalization: The new era of global flows,* McKinsey Global Institute.

Manyika, J., and CH. Roxburgh (2011), "The great transformer: The impact of the internet on economic growth and prosperity," *McKinsey Global Institute,* 1.

Mayzlin, D.; Dover, Y., and J. Chevalier (2014), "Promotional reviews: An empirical investigation of online review manipulation," *The American Economic Review,* 104(8): 2421–2455.

Milgrom, P., and J. Roberts (1982), "Predation, reputation, and entry deterrence," *Journal of Economic Theory,* 27(2): 280–312.

Nosko, CH., and S. Tadelis (2015), The limits of reputation in platform markets: An empirical analysis and field experiment, *Technical report,* National Bureau of Economic Research.

Resnick, P.; Zeckhauser, R.; Swanson, J., and K. Lockwood (2006), The value of reputation on ebay: A controlled experiment, *Experimental Economics,* 9(2): 79–101.

Rossi, M. (2018), *Reputation for what? moral hazard, adverse selection and the airbnb review system,* unpublished manuscript.

Sundararajan, A. (2016), *The sharing economy: The end of employment and the rise of crowd-based capitalism,* Mit Press.

Tadelis, S. (2016), "Reputation and feedback systems in online platform markets," *Annual Review of Economics,* 8: 321–340.

Vial, B., and F. Zurita (2017), "Entrants'reputation and industry dynamics," *International Economic Review,* 58(2): 529–559.

Zervas, G.; Proserpio, D., and J. Byers (2015), *A first look at online reputation on airbnb, where every stay is above average.*

INSIDE THE ENGINE ROOM OF DIGITAL PLATFORMS: REVIEWS, RATINGS, AND RECOMMENDATIONS[1]

Paul BELLEFLAMME[2]

Martin PEITZ

Abstract

The rise and success of digital platforms (such as Airbnb, Amazon, Booking, Expedia, Ebay, and Uber) rely, to a large extent, on their ability to address two major issues. First, to effectively facilitate transactions, platforms need to resolve the problem of trust in the implicit or explicit promises made by the counterparties; they post reviews and ratings to pursue this objective. Second, as platforms operate in marketplaces where information is abundant, they may guide their users towards the transactions that these users may have an interest in; recommender systems are meant to play this role. In this article, we elaborate on review, rating, and recommender systems. In particular, we examine how these systems generate network effects on platforms.

Keywords: Platforms, network effects, ratings, recommender systems, digital economics.

JEL classification: L80.

[1] We thank Juanjo Ganuza, Gerard Llobet, and Markus Reisinger for helpful comments. Martin Peitz gratefully acknowledges financial support from Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224. Paul Belleflamme acknowledges financial support from Aix-Marseille School of Economics and Kedge Business School.

[2] This work was prepared when Paul Belleflamme was employed by Aix-Marseille University and Kedge Business School whose support is gratefully acknowledged.

I. INTRODUCTION

Platforms can be defined as undertakings whose core mission is to enable and to generate value from interactions between users. Although platforms can operate off-line, Internet and digital technologies greatly contribute to reducing transaction costs, which explains why digital platforms are so prevalent nowadays. Digital platforms typically provide a number of services that generate so-called "platform-specific network effects," insofar as the attractiveness of a particular platform increases with the volume of interactions that the platform manages. Roughly speaking, the platform becomes more attractive the more it is used, and, as a result, each user cares about the participation of other users.[3]

The participation of other users may matter for a few reasons. First, their active evaluation of products and services, or the information contained in their actions, provides guidance for a user's action; second, the information contained in the users' actions enables the platform to provide better services or add specific offerings, both of which potentially benefit all users. In this article,[4] we focus on the former reason and analyze platforms' deployment of review, rating, and recommender systems. These non-price strategies allow platforms to generate within-group and/or cross-group external effects, that are (as we will argue below) platform-specific: the disclosure, aggregation and interpretation of information provided by the participants steer trade on the platform, thereby affecting the overall attractiveness of participating on the platform.

How are rating and recommender systems instrumental in producing network effects? Consider, for instance, the case of Amazon, which publishes product reviews and average ratings. Arguably, the more consumers that are active on Amazon, the more informative are the reviews and ratings, thus allowing consumers to make a better-informed decision. Amazon also provides recommendations by matching product descriptions with consumers' interests. Similarly, the more consumers that are active on the platform and the larger the volume of transactions they generate, the better the data that Amazon has about consumer characteristics and, so, the better the matches it can suggest; the quality of recommendations increases thus with the number of consumers, which in many cases will lead to a higher expected net consumer benefit. These mechanisms point to positive within-group external effects.

On two-sided platforms, positive cross-group external effects might arise. For instance, a high-quality seller thinking of participating on Ebay, Amazon

[3] For a justication of this broad notion of what constitutes a platform (*i.e.,* a managed marketplace featuring network effects), see, for instance, Belleflamme and Peitz (2018b).

[4] We use material from Chapters 2 and 5 of Belleflamme and Peitz (2018a).

Marketplace or some other B2C platform cares about the ease with which it can build its reputation. The more buyers active on the platform, the more precise the information about the seller type at a given point in time (assuming truthful consumer ratings). Thus, there is a positive cross-group external effect from buyers to high-quality sellers. Similarly, the more buyers on a platform, the better the matching between buyers and sellers (in terms of horizontal characteristics). This, in particular, reduces the expected number of products returned to the sellers. Thus, thanks to the recommender system, there is a positive cross-group external effect from buyers to sellers. This effect is strengthened by more detailed data on each consumer, as this improves the expected match quality.

Ratings are intended to help consumers make choices based on the quality or value-for-money dimension. Recommendations can also serve this purpose; they also have the potential to address buyer heterogeneity if they are personalized. This does not mean that some degree of personalization is impossible in the context of a rating system. In fact, several platforms offer the option of personalization; by, for instance, showing ratings and reviews only of buyers with certain profiles. Such rating selection can provide better guidance because what is good for one group of buyers is not necessarily good for others. For example, a business traveler may have different needs and preferences than a family on vacation and, thus, may prefer to see only reviews and ratings by fellow business travelers.

In the rest of this article, we analyze the economics behind the ratings, reviews and recommendations that have become mainstream on digital platforms. We start in Section II with rating and review systems. These systems provide platform users with information about either products or their counterparties to a transaction. Of crucial importance is, of course, the informativeness of these systems, which depends not only on the users' actions but also on the specific design chosen by the platforms. We then turn, in Section III, to recommender systems, which aim to reduce users' search cost by pointing them towards transactions that may better match their tastes. Besides the ability of such systems to generate network effects, we also discuss their effects on the distribution of sales between 'mass-market' and 'niche' products, as well as the incentives that platforms may have to distort their informativeness. We conclude in Section IV.

II. RATINGS AND REVIEWS

Ratings and reviews are prevalent on digital platforms. Platforms acting as vertically integrated retailers (such as Amazon.com) generally ask buyers to rate products or services and often give buyers the chance to write reviews. In such a

case, we speak of product ratings and *product reviews.* For platforms that host buyers and sellers (such as Amazon Marketplace), users on either side are often asked to rate and comment on the counterparty to the transaction. These we call seller (or buyer) ratings and reviews.

1. Asymmetric Information and Network Effects

Before analyzing the economics of rating and review systems, we consider their significance for digital platforms. Unquestionably, the main function of ratings and reviews is to respond to asymmetric information problems. At the same time, they are also an important source of network effects, which makes them instrumental in platforms' efforts to gain market shares. We describe these two aspects in turn.

1.1. Asymmetric Information

Asymmetric information problems are prominent on platforms that facilitate the trade of experience goods, as buyers typically have less information than sellers about the quality of the goods or services offered for sale. In this section, we focus on those asymmetric information problems that arise with experience goods.[5]

A traditional instrument to address asymmetric information problems is the use of *certification and warranties.* When a seller wants to transact with a buyer, third parties may provide certification, and platforms are a natural candidate for such certification services. Certification is an ex ante solution to asymmetric information problems, as it may ensure a minimum quality provided on the platform; lower-quality sellers are not admitted or worse-performing sellers are expelled from the platform. Certification can be mandatory or voluntary. For instance, Uber checks the records of its drivers to make sure that they are eligible to drive; such certification is mandatory. Airbnb offers the sellers of accommodation services the option to certify the authenticity of photos of the announced property, thus reducing the risk of unpleasant surprises for the buyer; such certification is voluntary. As for warranties, they may, in principle, be provided by sellers themselves, but platforms are often in a better position to provide them, since they interact more frequently and directly with buyers.

[5] We argue in Section III that asymmetric information problems may also apply to search goods. In this case, even if buyers can ascertain quality before purchase, they may lack information prior to investing time and effort to obtain relevant product information. Here, platforms can use ratings and reviews (on top of other instruments) to lower buyers' search costs and to improve the match between buyers and products/sellers.

Asymmetric information problems can also be addressed ex post through *insurance and guarantees.* For instance, Airbnb insures sellers against vandalism by buyers. Another example is Ebay's guarantee to buyers (introduced in 2010) to compensate them if the seller does not deliver as advertised (see Hui *et al.,* 2016).

Rating and review systems complement these classic instruments and tend to become relatively more effective than them, the larger the number of transactions that the platforms facilitate. Indeed, the ability of rating and review systems to tackle information problems faced by buyers (and possibly sellers) increases with the volume, variety, and velocity of the data that platforms can collect about their users and the transactions they conduct.[6]

1.2. Network Effects

As just argued, ratings and reviews can be an important source of network effects: the more users that are active on a platform –and, thus, the more ratings and reviews that are available– the better-informed other users are prior to making their purchase decisions. In the following sections, we will clearly identify the various forms that these network effects can take. What we want to stress here is that, although users often have access to ratings and reviews whether or not they purchase on a particular platform, network effects tend to be 'platform-specific' for a number of reasons.

First, some users may not consider purchasing on a platform different from the one on which they obtain information. In this case, even if a featured product is available on multiple platforms, it matters on which platform better information is available. For instance, in the early 2000s, buyers in the U.S. may have accessed ratings and reviews available on books at Amazon and then purchased the book from Barnes & Noble. However, as we discuss below, the positive sales effect of high ratings is more pronounced on the same platform than across platforms. This suggests that a substantial fraction of buyers only took note of reviews and ratings only on the platform on which they terminated their purchase.

Second, when buyers rate sellers on a two-sided platform, a seller may (at least partially) condition its behavior on the distribution channel picked by the user. In this case, the seller's reputation is actually conditional on the transaction on a platform. For example, a hotel may be more accommodating to the wishes and requests of a guest who booked on a particular platform. To give another example, a seller may exert particular effort to speedy delivery of a product ordered through a particular platform.

[6] The veracity of the data is also crucial, as we discuss in point 4 of Section II.

Third, the identity of a seller may be platform-specific, or it may be costly for the user to identify the same seller across platforms. For instance, it may be difficult to verify that the seller name on Ebay or Amazon Marketplace corresponds to the seller name on some other distribution channel. If this is the case, network effects are, by construction, platform-specific. For all these reasons, we can safely record the following finding.

Finding 1. Because they generate platform-specific network effects, rating and review systems fuel self-reinforcing mechanisms that, other things being equal, make successful platforms even more successful, at the expense of their smaller rivals.

We now turn to an in-depth analysis of rating and review systems on products and services (point 2 of Section II), and on transaction counterparties (point 3 of Section II). We then address the fundamental issue of the informativeness of these systems (point 4 of Section II).

2. Product Rating and Review Systems

Many online retailers have established rating and review systems (or 'rating systems' for short) that allow buyers to rate and comment on particular products. Absent such a rating system, we would not classify an online retailer as a platform, since, given prices, a buyer's purchase intention would not be affected by other buyers' purchases. However, the presence of a rating system renders the retailer a platform, as it is a source of network effects, and its design affects the strength of network effects.

Finding 2. Product rating systems have the potential to solve asymmetric information problems. In an e-commerce context in which buyers rate products, as more buyers on a platform make the average product rating more informative, a platform with a product rating system features positive network effects among buyers.

To illustrate this point, we consider a firm that carries products sourced at marginal cost c and sold at price p. Neither the firm nor the buyers know the quality of any product prior to consumption. What is known is that quality q may be either high ($q=H$) or low ($q=L$) with probability 1/2, and that this probability is drawn independently across products. Buyer valuations for high and low quality (respectively, v_H and v_L) satisfy $v_H > c > v_L$ and $(v_H + v_L)/2 > c$. The first set of inequalities tells us that if information were complete, only high-quality products would be traded (as buyers value the low quality below its marginal cost). The second inequality tells us that when buyers are

uninformed, trade will nevertheless take place, as the average valuation of a product is above the marginal cost.

Suppose that there are k buyers, who arrive in random order at each product. Each buyer is inclined to leave a review (if the firm provides a rating system) with some probability ρ, which is independent of the actual quality of a product. Furthermore, suppose that buyers perfectly observe product quality after purchase and report this quality truthfully if they write a review.

Absent a product rating system, a monopoly firm sets its price equal to the average valuation, $p = (v_H + v_L)/2$, and all buyers make a purchase. With a product rating system and under the assumption of a uniform price, the firm has to set the price such that buyers buy the product even when no review is available. This price is the same as without a rating system, as a buyer who does not observe any review is willing to pay up to the average valuation–*i.e.,* $(v_H + v_L)/2$.

At such a price, a buyer buys the product as long as no review of low quality has been posted (*i.e.,* if either no review is available, or if only positive reviews are available). If the product is of high quality, regardless of the order in which buyers appear, there will be no negative review posted. If the product is of low quality, a buyer in position k encounters with probability $(1-\rho)^{k-1}$ that none of the previous $k-1$ buyers left a review. Thus, the overall probability that a buyer in a market with a total of n_b buyers does not see a negative review is P_H+P_L, where $P_H=1/2$ is the probability that the product is of high quality (and it does not matter then whether or not buyers wrote a review), and $P_L=\sum_{k=0}^{n_b-1}(1-\rho)^k/(2n_b)=\left[1-(1-\rho)^{n_b}\right]/(2\rho n_b)$ is the cumulative probability that none of the previous buyers left a review and the product is of low quality. Importantly, P_L decreases as the number of buyers, n_b, increases (it converges to 0 as n_b turns to infinity). The expected surplus of a buyer is then equal to $U^e=P_H(v_H-p)+P_L(v_L-p)$. As $p=(v_H+v_L)/2>v_L$, it follows that $U^e=(P_H-P_L)(v_H-v_L)/2$, which is increasing in n_b. Thus, a platform with a product rating system is more informative the larger the number of buyers and, therefore, exhibits positive network effects.[7]

In the above example, the rating system generates positive network effects among buyers; such effects are generally called 'within-group' or 'one-sided' network effects. Does this imply that retailers with a rating system do not feature two-sidedness? In general, one- or two-sidedness is often a matter of

[7] In the example, a monopoly firm makes a lower profit with a rating system because it sells at the same price to fewer buyers. However, if buyer participation necessitates an up-front fixed cost for buyers, there is a hold-up problem absent a rating system. In this case, establishing a rating system limits the hold-up problem and, in equilibrium, may lead to higher profits for a firm with a rating system, since the market breaks down absent a rating system. In this case, a monopoly firm has the incentive to establish a rating system.

the concrete circumstances. This is also the case with rating systems, as we now show in the following three examples.

In the first example, we consider a stylized two-period setting in which some users simultaneously make purchase decisions in period 1, and other users simultaneously make purchase decisions in period 2. Suppose that a fraction of the former group posts a rating. Thus, period-2 buyers can make better-informed decisions, as the number of period-1 users increases. This means that due to the ratings system, there are positive cross-group external effects from period-1 users to period-2 users.

In the second example, we consider another stylized setting that features two types of buyers. For the first type, products are experience goods (quality is observed with some noise after purchase) and for the second type, they are credence goods (quality is not observed, even after consumption). Suppose that only users who learn the quality of the product rate the product (truthfully) and that those who do not learn the quality do not leave a rating. If users buy different products over time and base their decisions on average ratings, they benefit from a retailer attracting more type-1 buyers, as additional rankings allow for better-informed choices. Thus, there exist positive within-group external effects for type-1 buyers and positive cross-group external effects from type-1 to type-2 buyers. To the extent that type-1 buyers can draw on their own previous experience, informative ratings are less essential than for type-2 buyers, and, thus, the cross-group external effects generated by type-1 buyers are stronger than their within-group external effects.

Turning to the third example, consider now that, depending on the group a buyer belongs to, she leaves reviews with different probabilities; let λ_j denote the review probability in group j. If n_j^i buyers of group j participate on platform i, the expected number of reviews on platform i is $m^i = \lambda_1 n_1^i + \lambda_2 n_2^i$. More reviews make a platform more attractive to buyers. This benefit can be captured by an increasing and concave function $f(m^i)$. In this setting, there are positive within-group external effects for each group of buyers. In addition, there are positive cross-group external effects between the two groups of different strength (if $\lambda_1 \neq \lambda_2$).

As argued above, rating systems help buyers make more-informed choices. With a rating system in place, the empirical prediction is that a more-highly-rated product should see its sales increase compared to a less-highly-rated product. Chevalier and Mayzlin (2006) analyze the effect of book reviews on the sales patterns of the two leading online booksellers in the USA (at that point in time), Amazon and Barnes & Noble.[8] Both offer buyers the opportunity to post book

[8] Our exposition is almost identical to that in Belleflamme and Peitz (2015: Chapter 15).

reviews on their site. The central question of the study is whether an additional negative report on Amazon leads to a decline in sales at Amazon relative to the sales at Barnes & Noble. If the answer is 'yes,' this means that book reviews carry relevant information that affect sales. To answer this question, Chevalier and Mayzlin use the 'differences-in-differences' approach–that is, they take differences between the relative sales of a book at the two retailers to control for possible effects of unobserved book characteristics on book sales and reviews. Data were publicly available: they cover a random selection of book titles with certain characteristics in three short periods–two-day periods in May and August 2003 and May 2004.

Chevalier and Mayzlin regress the natural logarithm of the sales rank of book i at retailer j (which serves as a proxy for sales) on a number of variables including fixed effects, prices at Amazon and Barnes & Nobles and the share of positive (5-star) and negative (1-star) reviews. Chevalier and Mayzlin show that an additional positive review for a particular book at one retailer leads to an increase in the sales of this book at that retailer relative to the other. There is also some evidence that an additional negative review is more powerful in decreasing book sales than an additional positive review is in increasing sales (measured by the sales rank). The fact that the length of reviews also matters suggests that buyers not only use summary statistics but actually take a look at the reviews; this also suggests that they take the content of the review explicitly into account (perhaps to evaluate how much to trust a particular review or because there is uncertainty with respect to the fit of the match, which is buyer-specific).

Vana and Lambrecht (2018) use product review data from an UK online retailer. They identify the effect of the content of individual reviews, since the position at which reviews are placed is exogenous in their setting (placement by the date of being posted). When a new review appears, all existing reviews are shifted downward by one position. This shift occurs regardless of the content and rating of any review. As the authors show, the rating of the first displayed reviews have a strong effect of purchase likelihood. In particular, if these reviews come with a high rating (four or five stars out of five) the estimated purchase probability increases significantly.

3. Seller Rating Systems

So far, we have considered rating systems by a retailer that interacts with consumers. We now turn to rating systems of two-sided platforms: B2C and C2C platforms bring sellers and buyers together. Here, rating systems are a

solution to the general trust problems encountered by buyers. Should they trust the quality claims that sellers make about their products on offer? Should they trust the service promises? Possibly, these trust problems also exist the other way round. In a bilateral relationship, such trust problems can be solved through repeated interaction. When buyers are likely to provide reviews and/or ratings and these are informative, the trust problem can (at least, partially) also be solved in anonymous markets. Here, the rating and review system (or 'reputation system') serves as a substitute for personal experience: an individual buyer can draw on the collective experience of other buyers.

If you have ever booked a room in a hotel and learned upon late arrival that all the rooms were occupied, you may appreciate booking platforms that provide feedback from other buyers on the reliability of the information provided by the hotel. Perhaps more importantly, hotels have to worry about their reputation if they do not treat their guests well. For this reason, reputation systems are an important driver of the success of platforms as enablers to transaction –they may generate trust for at least one of the parties involved and resolve asymmetric information problems.

A rating system may be one-sided or two-sided. For instance, Amazon Marketplace has a one-sided rating system according to which buyers rate sellers. The initial Ebay system was two-sided, and so are the systems of Airbnb and Uber. Here, each transaction partner can rate, and leave a review about, the partner on the other side.

Rating systems can tackle adverse selection and moral hazard problems. For instance, accommodations on Airbnb that suffer from some unexpected problems can be singled out by reviews and ratings. To the extent that these unexpected problems are inherent to the property, this reveals the quality of the accommodation and resolves adverse selection problems. Unexpected problems can also arise if the seller does not exert effort; here, ratings and reviews can help to solve the associated moral hazard problem.

If reviews and ratings are noisy, a platform with few transactions per seller does not provide very reliable information. Given the number of sellers, the more buyers that are active on the platform, the more precise is the information on any seller since the average valuation tends to converge to the true valuation. This suggests that there exist positive network effects on the buyer side–we will discuss and qualify this finding below (as the informativeness of the ratings depends on their truthfulness).

Finding 3. Seller rating systems have the potential to solve asymmetric information problems. In a buyer-seller context in which buyers rate sellers, as more buyers on a platform make the rating system more informative, a platform with a rating system features positive within-group external effects on the buyer side.

For a given number of buyers, the rating system's informativeness tends to increase in the response rate of buyers. Here, the rating system may be designed to encourage buyers to leave a review or rating. Response rates may depend positively on the ease of use of the platform, and on the community feeling that it creates. The platform may also provide non-monetary or monetary incentives to leave reviews. As an example of the former, Tripadvisor awards a number of badges depending on review activity. Regarding the latter, Fradkin, Grewal, and Holtz (2017) ran a field experiment on Airbnb in which they provided monetary incentives for leaving reviews and showed that this can be effective. A seller reputation system may also suffer from low response rates by buyers who are afraid to rate a seller after a bad experience –more on this below when we discuss the informativeness of ratings and reviews.

A number of empirical works have shown that more reputable sellers are more successful –that is, reputation pays. Reputable sellers may be able to ask for a premium and/or they may enjoy higher transaction volumes– in particular, they may also be able to successfully sell products that buyers *a priori* deem to be risky to buy.

Resnick *et al.* (2006) run a controlled field experiment to investigate the price premium of reputation: they sell a number of identical products (collectible postcards); some of them are randomly assigned to an established seller with a good record and some to a seller with little track record. They estimate an 8% price premium for a seller with 2,000 positive and one negative ratings, compared to a seller with ten positive and zero negative ones. Cabral and Hortacsu (2010) collect a large data set of seller histories on Ebay. Unfortunately, they do not observe the number of a seller's past completed transactions and assume that the frequency of a seller's feedback is a good proxy for the frequency of actual transactions.[9] According to their estimates, a seller's weekly sales growth rate drops from a positive rate of 5% to a negative rate of 8% upon receiving his first negative rating.[10]

---

[9] This assumption may seem innocuous. However, as discussed below, different seller types are likely to have different rates by which buyers give reviews and ratings.

[10] A potential drawback is that they do not include price effects, but they may actually be small. Other early empirical work on auction sites includes McDonald and Slawson (2002), Melnik and Alm (2002), Linvingston (2005), and Jin and Kato (2006). For a summary of this and other work, see Bajari and Hortacsu (2004) and Tadelis (2016).

Some platforms started off without a rating system. For instance, the Chinese auction site Eachnet operated initially (1999-2001) without such a system. A certain degree of bilateral trust between seller and buyer was established through communication between the two parties, which eventually led to a physical meeting. Thus, the buyer could inspect the product before paying, and the seller could make sure that the seller made the payment. While this does not resolve all asymmetric information problems ex ante, some of the most unpleasant surprises for buyer and seller could be avoided even without a rating and review system. In 2001, Eachnet introduced a rating and review system. Cai *et al.* (2014) empirically investigate how a seller's "reputation" affects outcome, depending on whether a rating and review system is in place. A seller's reputation is approximated by the cumulative success rate of its listings. A seller's listing is successful if it led to at least one transaction. One may expect that if a buyer and a seller successfully complete a transaction, they may be more likely to interact again in the future. This may hold, in particular, for "reputed" sellers (*i.e.,* those with a high cumulative success rate). Indeed Cai *et al.* (2014) find a positive correlation between sellers' cumulative success rate and the fraction of repeat buyers. The important finding here is that this correlation weakens after the introduction of the rating system. This suggests that the rating system makes the asymmetric information problem faced by occasional buyers less severe and, thus, serves as a partial substitute to reputation within a bilateral relationship.

The introduction or redesign of a rating system may have an impact on the sellers' decision of whether to join a platform (and on the scale of its activities). For instance, if the rating system leads to better-informed buyers, low-quality sellers may abstain from participating. It might also affect the behavior of sellers beyond whether (and with what intensity) to participate. For instance, if a misrepresentation of product quality is punished through a negative rating that is easily observable to potential buyers, a seller may be more careful in drafting his announcements. In short, a rating system may affect participation (and, thus, affect the amount of adverse selection) and behavior, given participation (and, thus, the degree to which the moral hazard problem plays out). Klein, Lambertz, and Stahl (2016) investigate the effects of Ebay's redesign of its reputation system in May 2008, when Ebay introduced one-sided feedback that is not subject to retaliation and, thus, can be seen as more accurately reflecting a buyer's experience (below, see more on retaliation). Since, prior to that date, in May 2007, Ebay introduced an anonymous details seller rating (DSR) on top of its rating system, Klein, Lambertz, and Stahl could use this DSR before and after the change to a one-sided rating system as a measure of buyer satisfaction. They found a significant increase in buyer satisfaction with the introduction of the one-sided rating system, but did not observe a significant change in the sellers' exit rate. This can be seen as evidence that, in this instance, the redesign of the rating system was successful in reducing moral

hazard but did not significantly affect the composition of sellers. In the case of Ebay, this seems conceivable, as a low-quality product may find its buyer even if quality is revealed since there may be a market for such low-quality products. The effect of the redesign of the rating system would then encourage truthful announcements by sellers but would not remove their incentive to participate.

Finding 4. In the case of hidden-information problems, sellers are affected differentially by seller rating systems: high-quality sellers enjoy a positive cross-group external effect from more buyers leaving ratings, while low-quality sellers suffer a negative cross-group external effect from more buyers leaving ratings. In the case of hidden-action problems, all sellers may benefit, as buyers understand that the system disciplines sellers.

4. The Informativeness of Ratings and Reviews

Ratings and reviews can be relevant for buyers only if they contain relevant information. Clearly, if they are informative about the (price-adjusted) quality of a product, buyers must, at least to some degree, have a common perception of the (price-adjusted) quality, and buyers must be able and willing to report their experiences with the product.

We identify three sets of reasons why the informativeness of ratings and reviews may be limited due to decisions by buyers and sellers[11]: (i) noisy ratings and reviews; (ii) strategically distorted ratings and reviews; and (iii) asymmetric herding behavior. We discuss these, in turn, before examining how platforms can act to make rating systems more–or less–informative.

4.1. Noise

We describe here four reasons that buyers may leave noisy ratings and reviews: bad understanding, idiosyncratic tastes, uncontrollable shocks, and price variations.

— Bad understanding

Buyers may leave noisy ratings and reviews simply because they fail to understand what they are asked. While this is often easily identified after reading a review, buyers who rely on summary statistics may not be able to identify that ratings are based on irrelevant experiences. For instance, this applies to

[11] For other overviews, see Aral (2014) and Tadelis (2016).

product ratings on Amazon. Here, some reviewers do not base their rating on the quality and characteristics of the product they bought, but on such factors as Amazon's delivery service, which can be considered orthogonal to the product sold by Amazon. For example, the 2010 edition of our textbook *Industrial Organization: Markets and Strategies,* received a 5-star rating by one reviewer on Amazon.com with the following review: "It's my first time to buy used books. And it has definitely met my expectation. Well kept just few marks. Like it very much."[12] While we are happy that the reviewer gave a 5-star rating, we are not so sure if this actually reflects his or her quality assessment of the book rather than the physical appearance of the used copy.

— *Idiosyncratic tastes*

Ratings may also be noisy for potential buyers because of idiosyncratic tastes. While rating systems are supposed to capture the quality of a product or seller, reviewers may comment on horizontal characteristics or on vertical characteristics for which they have heterogeneous willingness to pay. In other words, ratings that aggregate tastes of other buyers may not strongly correlate with one's own taste. For instance, a reviewer may give a negative product rating because she does not like the color of the product, but other potential buyers may not share this negative feeling.

— *Uncontrollable shocks*

Relatedly, there may be shocks that are not under the seller's control. If a reviewer leaves a negative seller rating because of late delivery, this may not have been under the seller's control if, say, the transport company did not deliver in time. One would expect that such shocks to product and service satisfaction wash out if there is a large number of reviewers. Thus, the informativeness increases with the number of fellow users, a source of the network effects mentioned above.

— *Price variations*

Product and seller reviews are often likely to be based on how satisfied a buyer is when taking into account how much she paid. However, products may be sold at different prices over time and space. Thus, what looks like a rather bad deal at a high price may be a good deal at a low price. Therefore, with price variation (over time and space), the informativeness of ratings suffers.

[12] As Tadelis (2016, p. 328) notes, confusion is likely with multiple review targets: "Multiple review targets may create an inference problem that confuses between the seller's quality of executing the sale and the quality of the product."

4.2. Strategic Distortions by Buyers or Sellers

Buyers or sellers may take actions that systematically distort seller or product ratings. Clearly, since sellers benefit from a positive reputation, they may pay others to leave positive reviews and ratings about their offers; they may also pay others to leave negative reviews about the offers of close competitors. First, we examine such 'fake reviews,' and then we consider the specific problems that may emerge from 'two-sided rating systems,' in which both counterparties to a transaction are invited to rate one another.

— Fake reviews

The unsuspecting reader may think that fake reviews are an issue cooked up by economists who believe in incentive theory. However, there is evidence that fake reviews are widespread and that markets for such fake reviews have been created (see, *e.g.,* Xu, Chen and Winston, 2015).[13]

Generating such fake reviews is costly. Costs and benefits from fake reviews depend on the particular site. As Ott, Cardie, and Hancock (2012) argue in case of hotels, the costs of a fake review are high if a user is required to purchase a product prior to reviewing it. For instance, hotel booking platforms Booking and Expedia require an actual purchase, whereas Tripadvisor (which, as a referral website, does not monitor transactions) allows anyone who claims to have made a booking to post reviews about a hotel. Thus, fake reviews are more costly on Booking and Expedia than on Tripadvisor. The expected benefit depends on the attention that a particular review attracts. Everything else being given, the benefit on a website with many visitors is greater, while on a website with many other reviews the expected benefit, it is smaller. Hence, in an environment in which the ratio of reviews to traffic is the same across websites, it is not clear on which website the expected benefit is the largest. We note that posting a fake review on a website with a quickly growing visitor base and a small stock of reviews is particularly attractive. This suggests that newcomer platforms must think hard about how to design their rating system right from the start.

Providing evidence on the extent of fake reviews is hard, since actual fakes are difficult to spot. Mayzlin, Dover, and Chevalier (2014) exploit different policies by hotel information and booking sites about who can leave feedback: Expedia requires the reviewer to have booked a hotel on its site, while Tripadvisor does not (as it only referred to booking sites). Thus, we would expect to see

[13] Since fake reviews are costly to generate, a more benign view of the use of positive, paid-for reviews and ratings is that they can be seen as a seller's costly advertising and may be used as a signal of high quality– the seminal paper on advertising as a quality signal is Milgrom and Roberts (1986). For an empirical analysis of such behavior on the platform Taobao, see Li, Tadelis, and Zhou (2016).

more fake reviews on Tripadvisor. Consider a geographic area in which hotels compete for business travelers. It is in the strategic interest of any hotel in this area to improve its ranking relative to that of competing hotels in the same area. A hotel can achieve this by inflating its own rating with fake positive reviews and by deflating the rating of hotels in its vicinity with fake negative reviews.

Mayzlin, Dover, and Chevalier argue that independent hotels are more likely to sponsor fake reviews, as their cost from being detected is less severe than if such a review was sponsored by a hotel belonging to a chain. Thus, the prediction is that hotels in the vicinity of such independent hotels have more negative reviews on Tripadvisor relative to Expedia, and independent hotels have more positive reviews on Tripadvisor relative to Expedia. These predictions are confirmed in their dataset. And fake reviews are not unique to hotels; for instance, Luca and Zervas (2016) analyze fake restaurant reviews on Yelp.

— *Two-sided rating systems*

Problems of systematic misrepresentation and, possibly, underreporting of negative experiences may arise with two-sided rating systems in which both buyer and seller leave feedback. Such two-sided ratings appear to be desirable if both parties have private information and/or choose private actions. In its early days, Ebay used a two-sided system, arguably because sellers would like to know which buyers to trust. In particular, a buyer may place the highest bid but then refuse to make the promised payment. With developments in electronic payments, this risk for the seller could be eliminated. This has removed the main reason to use two-sided ratings on Ebay. Other platforms continue to employ two-sided rating systems. This applies, in particular, to platforms in the sharing economy because here, not only the payment, but also the way a buyer uses a product matters to the seller. For instance, somebody renting out an apartment on Airbnb may worry about whether the renter will create a mess or damage some furniture.

Although two-sided rating systems do not necessarily distort ratings, the past system on Ebay did. The Ebay rating system had the design feature that buyers and sellers had a time window during which they could leave a feedback. When one party left a feedback, it was disclosed to the other party. This opened up the possibility of retaliation for a negative rating. Bolton, Greiner, and Ockenfels (2013) analyze rating behavior on the old Ebay and document that the two ratings in buyer-seller pairs are highly positively correlated. They also document that sellers typically wait for the buyer to leave a rating and respond promptly. This supports the view that sellers use their feedback as an implicit threat to leave a negative rating if they receive a negative one. This makes it more painful for buyers to give negative ratings and, effectively, distorts the

distribution of ratings received by sellers.[14] Indeed, as Nosko and Tadelis (2015) report, using internal Ebay data, a buyer is three times more likely to complain to Ebay's customer service than to give a negative rating. This suggests a severe underreporting of negative experiences. As mentioned above, Ebay eventually switched to a one-sided rating system.

Airbnb also has a two-sided rating system.[15] Initially, reviews were immediately made public, allowing the possibility of retaliation. Fradkin, Grewal, and Holtz (2017) run field experiments and find that those who do not provide reviews tend to have worse experiences than those who do. They conclude that strategic reviewing behavior has occurred on Airbnb, although the overall bias appears to be small. Also, since buyer and seller may interact socially, they may be less inclined to leave negative reviews.

Airbnb no longer makes reviews public as long as the counterparty still has the option of posting a review and has not yet done so. While one party does not observe the counterparty's review prior to uploading her own review, there remain reasons for strategically underreporting negative experiences (in addition to the social interaction reason given above). Reviews are not anonymized, so somebody who rents out a flat can check the track record of somebody wanting to rent the flat. If that person tends to leave negative reviews, a future landlord may be less inclined to confirm the request. Anticipating this, the potential renter may be less harsh and leave positively biased reviews or no review at all.

A platform has various design options that affect the response rate and the informativeness of review and rating systems. For our purposes, we summarize the insights obtained so far by the following finding.

Finding 5. Rating systems may suffer from a lack of informativeness due to noise and bias introduced through the actions of buyers and sellers. In particular, platform users may game the system. This tends to reduce the strength of network effects.

4.3. Asymmetric Herding Behavior

A tendency to provide positive feedback, but to refrain from providing negative feedback, does not necessarily arise due to strategic considerations or independent mistakes by reviewers. It may also be the result of asymmetric

[14] There is, of course, an easy way for the platform to avoid such retaliation possibilities: ratings may be disclosed only after the other party has provided the rating, or the time window to leave ratings has closed.

[15] For descriptive statistics on Airbnb's rating system, see Zervas, Proserpio, and John (2015).

herding behavior. Muchnik, Aral, and Taylor (2013) conduct a randomized field experiment with fake ratings of comments on posted articles on a news website and analyze the dynamics of future feedback. They observe an asymmetric response to a fake positive rating compared to a fake negative rating. They find that a fake positive rating increases the probability of accumulating positive herding by 25%. While a fake negative rating also increases subsequent negative votes, this was neutralized by offsetting positive votes. Thus, there is herding on positive but not on negative ratings –Muchnik, Aral, and Taylor call this a 'social influence bias.'

These results were obtained in a news setting and not in shopping contexts, but they are suggestive of reviewer behavior also in the latter contexts. This suggests that paid-for fake positive reviews can generate positive herding on B2C and C2C platforms. Thus, the damage done from a positive fake review would not be corrected if the fake report were not removed immediately but at some later time (see Aral, 2014). As pointed out above, there are other reasons that ratings and reviews do not provide accurate information. This may also give rise to long-term effects thanks to herding.

4.4. Design of the Rating System

In the analysis above, we identified reasons that rankings and reviews lose informativeness because of the actions taken by the transaction partners. The assumption was that the platform aims to maximize the informativeness, possibly battling against errors and gaming. While more-informative rankings and reviews tend to make the platform more attractive (and are a source of positive network effects), a for-profit platform is ultimately interested in maximizing profit. It may, then, have an incentive to sacrifice informativeness if that increases its revenues. In addition to measures taken by the platform that affect the aggregate rankings of products or sellers, the platform may vary the ordering and display of individual reviews. The findings by Vana and Lambrecht (2018) provide some indications how a different design of the listing of reviews can affect purchase probability.

The literature on certifying intermediaries provides some insights into the design of rating systems by a profit-maximizing platform. In particular, platforms may deliberately design their system so as to avoid the worst offending behavior –that is, it features a minimum quality threshold– but to offer few clues about product quality otherwise. In such a case, rating inflation and presumed design flaws that limit the informativeness of a rating system would actually indicate that a profit-maximizing intermediary with market power sacrifices buyer participation in favor of higher margins. This is the lesson one can draw

from the work on certifying intermediaries by Lizzeri (1999), who shows in an adverse selection environment that a platform discloses only whether a product satisfies a minimum quality threshold.[16] In his setting, a monopoly intermediary charges a fee to sellers for providing its certification service.[17] As a result, the intermediary certifies minimum quality for products that are traded via the intermediary. Translated into the context of rating systems, the platform commits to its rating system and charges sellers for being listed. Thus, Lizzeri's result says that the rating system is designed in such a way that only the worst offenders disappear from the platform.

Finding 6. A profit-maximizing platform may deliberately design its rating system so as to limit its informativeness. As a result, sellers of rather low quality may do better on such a platform than on a platform that maximizes the quality of its rating system, while high-quality sellers do worse.

Bouvard and Levy (2016) further investigate the potential tension between informativeness and rent extraction. In their setting, the platform cannot commit to a certification technology and establishes a reputation for accuracy; for its service, it charges a fixed fee to participating sellers upfront. Applied to ratings systems, this means that the platform can redesign features that reflect the rating system's accuracy; and the fixed fee corresponds to a listing fee charged to sellers, as is observed, for example, on some price search engines.

Sellers have different opportunity costs of providing high quality. While higher accuracy attracts high-quality sellers, it repels low-quality sellers. As a result, the profit of a platform is first increasing and then decreasing in the level of accuracy it provides to sellers seeking certification. Thus, a profit-maximizing platform provides an intermediate level of accuracy. Applied to rating systems, instead of offering certification, a platform may make use of buyer reviews and ratings to (noisily) reveal quality. The design decisions regarding the rating system then affect its accuracy.

Platform competition improves the information available to buyers when sellers have to make a discrete choice between platforms: it enables full disclosure in the Lizzeri's (1999) setting and increases accuracy in Bouvard and Levy's

[16] Similarly, Albano and Lizzeri (2001) analyze a moral hazard problem.

[17] The timing is as follows: first, the intermediary sets its fee and commits to an information disclosure policy. Second, after observing the intermediary's decision, sellers decide whether to pay the fee, offer their products through the intermediary, and submit their product for testing. Third, consumers observe all previous decisions, and the seller makes a take-it-or-leave-it offer.

(2016) setting. By contrast, under seller multihoming, Bouvard and Levy (2016) show that platforms have weaker incentives for accuracy under competition.

III. RECOMMENDATIONS

As we discussed in the previous section, buyers can obtain valuable information from reviews and ratings by other buyers. In this case, the role of the platform is twofold: first, it invites buyers to evaluate various offers that have proved successful or popular with others; second, it organizes the exchange of the information across users (possibly combined with some policing so as to ensure that abuses are contained and mistakes are corrected). Since buyers actively provide and access the information, we may consider ratings and reviews as part of a platform's *information-pull* strategy.

In this section, we examine an alternative strategy of platforms, which consists of making recommendations to specific buyers. Such recommendations, based on popularity and on other sources of information, are an attempt to reduce search costs. Hence, platforms pursue an *information-push* strategy, as they advertise specific products to buyers based on their characteristics and observed behavior. Naturally, information-pull and –push strategies are not mutually exclusive– quite the contrary, as ratings and reviews often serve as inputs for recommendation algorithms. For instance, Amazon makes product suggestions, and buyers then access additional information before making their purchase decision.

In what follows, we first analyze how recommender systems, such as rating systems, generate network effects (point 1 of Section III). Next, we examine how recommender systems affect the distribution of sales (point 2 of Section III): do they contribute to making popular products even more popular, or do they drive consumers to discover niche products? Finally, we look into platforms' incentives to manipulate recommender systems (point 3 of Section III).

1. Product Recommender Systems and Network Effects

In this Section, we argue that product recommender systems are the source of positive network effects. This insight is easily established when buyers have homogeneous tastes and make mistakes, and the recommender system is based on the popularity of a product. Suppose that there are two products that can be ranked by their attractiveness. Product A is more attractive than product B; more specifically, suppose that product A gives a net benefit of 1 and product B of -1. Consumers arrive sequentially and can be of two types:

'amateur' or 'expert.' An amateur consumer bases her decision on popularity, while an expert consumer acquires information about product features and makes a purchase based on that information.

To construct a numerical example, suppose that 50% of buyers follow a recommendation if they receive one and otherwise do not buy, while the remaining 50% collect information and, with 80% probability, make the right choice–*i.e.,* with 20% probability, they erroneously choose the inferior product. The recommender system recommends the product that is purchased more. We will show that the last buyer is better off if there are more fellow buyers. Let us start with two buyers. If buyer 2 is an amateur, she makes an expected benefit $0.5(0.8-0.2)=0.3$, as, with 50% probability, buyer 1 was an expert (that is, buyer 1 purchased and, thus, indirectly recommended, the 'good' product with 80% probability and the 'bad' product with 20% probability). If buyer 2 is an expert, she makes an expected benefit of $0.8-0.2=0.6$. Hence, the expected benefit of buyer 2 is 0.45 (*i.e.,* the average of 0.3 and 0.6, as she has equal chances of being either type).

Now consider the case with three buyers. If the third buyer is an expert, her expected benefit continues to be 0.6 (as the recommender system has no influence on her decision). If the third buyer is an amateur, she purchases only if the recommender system points her to the most popular product. For this to happen, the two previous buyers must have purchased one product more than the other. Let us examine when this does and does not happen. Four cases have to be distinguished according to the type of the successive buyers; each case has the same probability of occurrence–25%. The first case is the succession of two amateurs: as neither of them purchased, the recommender system remains silent, and the third buyer does not purchase either, yielding her a benefit of zero. Second, if the first buyer is an amateur (who, therefore, did not purchase) and the second is an expert, then the system recommends the good product with an 80% probability, and the bad product with a 20% probability, yielding the third buyer an expected benefit of $0.8-0.2=0.6$. Third, if the first buyer is an expert and the second an amateur, the configuration is similar to the previous one (as the second buyer follows the recommendation resulting from the first buyer's purchase decision); the expected benefit of the third buyer is again equal to 0.6. Finally, if there is a succession of two experts, both must have made the same choice for the recommender system to be informative (and so for the third buyer to purchase); this is so if they both decide to buy the good product (with 64% probability) or the bad product (with 4% probability); the third buyer's benefit in this case is then equal to $0.64-0.04=0.6$. In sum, if the third buyer is an amateur, her expected benefit is $0.25\times0+3\times0.25\times0.6=0.45$. Hence, the expected benefit of the third buyer is $0.5\times0.6+0.5\times0.45=0.525$.

Comparing the two cases, we observe that the last of three buyers has a larger expected benefit (0.525) than the last of two buyers (0.45). Hence, we have established that the last buyer benefits if more previous buyers are around and that buyers, prior to knowing their position in the sequence, are also better off if more fellow buyers are present. In this example, amateurs benefit from more buyers, as it becomes more likely that an expert has been around previously.

Finding 7. By recommending more-popular products, product recommender systems have the potential to provide purchase-relevant information to amateur buyers. In an e-commerce context, they have the potential to generate network effects, as a buyer is better off the more fellow buyers that are around.

A recommender system may also help to reduce the search cost. Suppose that there are several products, some of which are considered clear failures and a few that can be considered serious options. Absent recommendations based on popularity, a consumer may have to inspect quite a large number of products. With such recommendations, the consumer can restrict her search to the subset of serious options and, thus, reduce her expected search costs.

Finding 8. Product recommender systems have the potential to reduce search costs. In an e-commerce context, they have the potential to generate network effects, as a larger number of buyers provides more reliable information about which products are serious options.

If some consumers are frequent shoppers, while others buy only occasionally, the former make larger contributions to the functioning of the recommender system than the latter. As an illustration, suppose that frequent shoppers buy several products from a large set, whereas occasional buyers buy only one. The shopping behavior of frequent buyers allows the recommendation system to help other frequent shoppers to more easily find other products of interest. Thus, the recommender system generates positive within-group external effects among frequent shoppers.

If the recommender system can access additional information on occasional shoppers (*e.g.,* that they are close to certain frequent shoppers in a friendship network), information gathered on frequent shoppers may also allow for useful recommendations to casual shoppers. In this case, there is a positive cross-group external effect from frequent shoppers to occasional shoppers. By contrast, information on purchase decisions by occasional shoppers is of little or

no help in making better recommendations to other shoppers. More generally, not only the total number of users, but the composition of the recommendation network, matter for the functioning of the recommender system.

Recommender systems can also be important on *two-sided platforms.* Here, the platform can make recommendations to both sides with the aims of reducing search costs and improving expected match quality. These recommendations may be based not only on observables of the two individual users on either side, but also on the behavior of other users on both sides.

Finding 9. Partner recommender systems have the potential to reduce search costs. In a two-group matching context, they have the potential to generate positive cross-group external effects, as more participation by one group generates the chance for the platform to propose matches that are more attractive for members of the other group, and vice versa.

We note that while both sides tend to benefit from such cross-group external effects, the benefits may vary depending on the terms of transaction between users on both sides. These terms of transaction for a particular user may also depend on participation levels on the same side. For instance, if buyers for collectibles receive better recommendations, they may drive up the price and, thus, receive a smaller fraction of the generated surplus.

2. Product Recommender Systems and the Long Tail

In many internet markets, a limited number of items (often a few hundred) account for the bulk of sales, while the vast majority of items (which constitute the tail of the distribution) sell only very few units. It has been argued that internet markets have a longer tail in the sales distribution than traditional markets.[18] The question we address in this section is how recommender systems affect the distribution of sales: do they reinforce the skewness of the distribution, or do they make the tail longer, or thicker? We first discuss the main effects that recommender systems can have; we then formalize the intuition in a specific model, before reviewing recent empirical work.

— Heterogeneous tastes and recommendations

Since buyers often do not have homogeneous tastes, a recommender system reporting the popularity of different products may provide information

[18] For an informal account, see Anderson (2006).

about which types of consumers may like a specific product. In particular, some buyers may be aware that they have a taste for niche products in a certain product category, whereas others may realize their preference for the standard products that cater to the taste of the mass market. Recommender systems may be based on popularity information–that is, information displaying in relative terms how often a product has been purchased. As a fictional example, consider a supermarket selling different types of cheese and providing popularity information. If you are new to the store and know that you like to avoid unpleasant surprises, you may opt for the popular cheese varieties. However, if you know that you like new taste experiences, you may opt for cheese varieties that are bought less frequently. In such a situation, the fact that a product has or has not been sold often provides valuable information to new buyers. A buyer with a niche taste may buy products that sold little in the past, whereas a buyer with a mass-market taste will purchase products that sold a lot in the past.

In practice, buyers may encounter products with mass or niche appeal and, in addition, suffer from not being able to judge product quality ex ante. It may then appear to be difficult to disentangle popularity information as a proxy for quality from popularity information as an indication of whether a product is a mass-market product–one that provides a good fit to the taste of many buyers– or a niche market product–one that provides a good fit to the taste of only few buyers.

There are two borderline cases. In the first, all buyers have the same taste and care only about quality. High quality proves to be more "popular" and accounts for a larger volume of sales if some consumers are informed about the product quality and buy only high quality, whereas others are not and, thus, have to randomize over several products of different qualities. Higher quality, then, turns out to be more popular. To resolve the asymmetric information problem, a platform may want to resort to a rating system, as analyzed in the previous section. Thus, the effect of such a rating system is to divert demand from a low-quality product to a high-quality product. In the other borderline case, buyers are uncertain only about whether the product better serves the mass or the niche market, leading to the outcome above.

A different situation arises if buyers observe whether a product is meant to cater to the mass or to the niche market, but they do not observe the product quality. To address the role of popularity information in guiding buyer behavior in such a situation, we present a simple model in which firm behavior is treated as exogenous –in particular, the prices of all products are fixed. As we will show, in such a scenario– in which consumers know in advance whether some product features fit their taste but are not fully informed about a quality dimension of

the product–a recommender system reporting the popularity of a product may also provide valuable information to consumers.

— A specific model

The model goes as follows.[19] Suppose that consumers face a choice problem of buying one unit of two products offered by two different sellers; they may buy none, one, or both. Prices are fixed throughout the analysis. With probability $\lambda > 1/2$, a consumer thinks more highly of product 1 than of product 2; consequently, product 1 can be called a mass-market product and product 2 a niche product. Each product can also be of high or low quality with equal probability.

The consumer's utility depends both on the quality of the product and on whether the product matches her taste. A high-quality product that provides the wrong match is assumed to give net utility $v_H = 1$ and a low-quality product, $v_L = 0$. A product with the right match gives the previous net utilities augmented by t. These utilities are gross of the opportunity cost z that a consumer incurs when visiting a seller (*e.g.,* clicking onto its website). A consumer knows her match value and receives a noisy private signal about quality. The noisy quality signal may come from noisy information in the public domain, such as publicly revealed tests. The ex ante probability of high quality is assumed to be 1/2. The probability that the signal provides the correct information is ρ, which, for the signal to be informative but noisy, lies between 1/2 and 1. Hence, with a positive signal realization, the posterior belief that the product is of high quality is ρ. It follows that if a consumer who prefers product i receives a high-quality signal and buys from seller j, she obtains expected utility $U_{Hg} \equiv \rho + t - z$ if $i = j$ (*i.e.,* if seller j offers the product that matches consumer i's taste), and $U_{Hb} \equiv \rho - z$ if $i \neq j$. Correspondingly, with a low-quality signal, expected utility is $U_{Lg} \equiv (1 - \rho) + t - z$ if $i = j$ and $U_{Lb} \equiv (1 - \rho) - z$ if $i \neq j$. Table 1 displays the four possible levels of expected utility.

TABLE 1

EXPECTED UTILITY ACCORDING TO SIGNAL AND MATCH

	Good match	*Bad match*
High-quality signal	$U_{Hg} \equiv \rho + t - z$	$U_{Hb} \equiv \rho - z$
Low-quality signal	$U_{Lg} \equiv (1 - \rho) + t - z$	$U_{Lb} \equiv (1 - \rho) - z$

[19] The model exposition is, in large part, identical to the one in Belleflamme and Peitz (2015: Chapter 15). It is based on Tucker and Zhang (2011).

For a given match, $\rho > 1/2$ implies that the consumer is better off with a high-quality signal: $U_{Hk} > U_{Bk}$ for $k = g, b$. Also, for a given signal, $t > 0$ implies that the consumer prefers to have a good match: $U_{Kg} > U_{Kb}$ for $K = H, L$. What is unclear is how the consumer balances the quality of the match with the quality of the signal. The consumer finds the quality of the match more important if $U_{Lg} > U_{Hb}$, which means that she is better off with a low-quality signal and a good match than with a high-quality signal and a bad match. This is so if $1 + t > 2\rho$. Otherwise, the quality of the signal outweighs the quality of the match.

We first consider the product choice of a single buyer–this is the situation encountered by buyers when no recommender system is available. A buyer purchases the product independently of the signal realization and match value if $U_{Lb} > 0$; that is, the opportunity cost of visiting a seller is sufficiently small, $z < z_{Lb} \equiv 1 - \rho$. By contrast, if the opportunity cost is too large, the consumer will never buy. This is the case if $U_{Hg} < 0$, or, equivalently, if $z > z_{Hg} \equiv \rho + t$. Hence, we focus on the intermediate range where $z \in [z_{Lb}, z_{Hg}]$. A product with a good match but a low-quality signal is bought if $U_{Lg} \geq 0$, or, equivalently, if $z \leq z_{Lg} \equiv 1 - \rho + t$. A product with a bad match but a high-quality signal is bought if $U_{Hb} \geq 0$ or $z \leq z_{Hb} \equiv \rho$.

As indicated above, two scenarios are possible. In the first scenario, the buyer sees the quality of the match as more important; the inequality $U_{Lg} > U_{Hb}$ is equivalent to $z_{Lg} > z_{Hb}$, which becomes $1 + t > 2\rho$. Thus, for this scenario to apply, consumer tastes must be sufficiently heterogeneous (t large) and signals sufficiently noisy (ρ small). In the second scenario, the quality of the signal matters more; we have $U_{Lg} < U_{Hb}$, or, equivalently, $z_{Lg} < z_{Hb}$. Thus, for this scenario to apply, consumer tastes must be sufficiently homogeneous (t small) and signals sufficiently informative (ρ large). Consumer choice can be fully described depending on whether $z_{Lg} > z_{Hb}$ or the reverse inequality holds.[20]

Second, we analyze buyer behavior in the presence of a *recommender system* that provides popularity information. For a recommender system to have any impact, we need at least another consumer who makes her choice after obtaining the information generated by the first consumer's choice. The

[20] For $z_{Lg} > z_{Hb}$, we obtain that a product is bought by a consumer who does not observe a low-quality signal and a bad match if $z \in (z_{Lb}, z_{Hb})$; it is bought by a consumer who observes a good match if $z \in (z_{Hb}, z_{Lg})$; and it is bought by a consumer who observes a good match and a high-quality signal if $z \in (z_{Lg}, z_{Hg})$. For $z_{Lg} < z_{Hb}$, we obtain that a product is bought by a consumer who observes neither a low-quality signal nor a bad match if $z \in (z_{Lb}, z_{Lg})$; it is bought by a consumer who does not observe a low-quality signal if $z \in (z_{Lg}, z_{Hb})$; and it is bought by a consumer who observes a good match and a high-quality signal if $z \in (z_{Hb}, z_{Hg})$. Interestingly, in the first scenario, if $z \in (z_{Hb}, z_{Lg})$, consumer choice is determined purely by the match quality, whereas in the second scenario, if $z \in (z_{Lg}, z_{Hb})$, consumer choice is determined purely by the signal realization.

recommender system here simply reports the choice of the first consumer. The second consumer knows the parameters of the model but neither the signal realization nor the type of the first consumer. We assume that all random variables are i.i.d. across consumers (concerning the quality signal, this is conditional on true quality).

To analyze whether a recommender system favors mass-market products or niche products, we consider two cases: $z \in (z_{Lb}, \min\{z_{Hb}, z_{Lg}\})$ and $z \in (\max\{z_{Hb}, z_{Lg}\}, z_{Hg})$.[21] The former case is characterized by a relatively low cost of visiting sellers. Here, a consumer who observes a good match with a particular product always visits the corresponding seller. The consumer visits the seller of the product with a bad match only in case of high-quality information. This implies that click and purchase data still contain some useful information for the second consumer. The second consumer knows whether she has a taste for the niche product or the mass-market product. Hence, if she has a taste for the niche product, she knows that it is unlikely that the first consumer had the same taste. Therefore, it is quite likely that the first consumer's visit or purchase was driven by a positive realization of the quality signal. The opposite reasoning applies to a consumer who has a taste for the mass-market product. Here, click and purchasing data are less informative, thus implying that sellers of niche products benefit more from information on visits or purchases.

In the latter case, in which $z \in (\max\{z_{Hb}, z_{Lg}\}, z_{Hg})$, information on a *lack* of visits or purchases hurts the seller of the mass-market product more. While niche sellers are at a disadvantage matching consumer tastes, this disadvantage becomes an asset when it comes to consumer inferences about product quality. It increases the benefit due to favorable popularity information and reduces the loss due to unfavorable popularity information.[22]

Tucker and Zhang (2011) provide support for this theory in a field experiment. A website that lists wedding service vendors switched from an

[21] In addition, there are two intermediary cases–that is, $z \in (z_{Hb}, z_{Lg})$ for $1+t>2\rho$ and $z \in (z_{Lg}, z_{Hb})$ for $1+t<2\rho$. In the first case, in which $z \in (z_{Hb}, z_{Lg})$, the first consumer's choice does not reveal anything about her private signal. Hence, the recommender system does not contain any valuable information for the second consumer. In the second case, where $z \in (z_{Lg}, z_{Hb})$, the first consumer's choice is determined solely by the signal realization. The second consumer will then use the information provided by the recommender system to update her beliefs: she updates her quality perception upwards if a particular product has been bought (purchase data) or if the seller has been visited (click data). This implies that a previous visit or purchase increases the chance of subsequent visits and purchases. Here, the recommender system favors the sale of high-quality products.

[22] An interesting question, which we do not analyze here, is the possibility of rational herding. This is a situation in which consumers ignore their private information and rely fully on the aggregate information provided by the system. This means that learning stops at some point. A seminal paper on rational herding is Banerjee (1992). Tucker and Zhang (2011) also address herding in the present context.

alphabetical listing to a popularity-based ranking in which offers are ranked by the number of clicks the vendor receives. The authors measure vendors when located in towns with a large population as having broad appeal and when located in small towns as having narrow appeal. Tucker and Zhang find strong evidence that narrow-appeal vendors receive more clicks than broad-appeal vendors when ranked similarly in the popularity-based ranking.

Finding 10. Product recommender systems reporting product popularity may affect mass-market and niche products differently. Given a similar ranking, niche products tend to do relatively better with such a recommender system.

A prominent mix of various recommender systems is in place at Amazon.com. Perhaps the most notable example (at least in product categories in which consumers do not search among product substitutes) is that, when listing a particular product, Amazon recommends other products that consumers have purchased together with the displayed product. The economics of such a recommender system are different from a system that merely reports the popularity of products. It allows consumers to discover products that serve similar tastes and, thus, is likely to produce good matches at low search costs. Such a recommender system is based on previous sales and appears to be particularly useful in consumer decision-making for products that enjoy complementary relationships. It implies that products with no or limited sales will receive little attention. This reasoning suggests that recommender systems may work against the long tail, an argument in contrast to the view that people discover better matches on recommender systems. The latter view is based on the observation that consumers with very special tastes more easily find products that provide a good match to their tastes, so that they do not need to resort to very popular products or buy at random.

However, these two views are not necessarily contradictory. While the long-tail story refers to the diversity of aggregate sales, the discovery of better matches refers to diversity at the individual level. It might well be the case that people discover better matches through recommender systems but that they discover products that are already rather popular among the whole population. Hence, sales data in the presence of recommender systems may show more concentration at the aggregate level.[23]

[23] This point is made in the numerical analyses of Fleder and Hosanagar (2009). However, in their model, the recommendation network essentially provides information about the popularity of a product and does not allow for more fine-tuned recommendations.

— Empirical work on recommender systems

While the previous discussion brings interesting insights, empirical analyses will have to show whether recommender systems, indeed, lead to more concentrated sales; or whether the directed search, which is inherent in recommender systems, reduces users' search costs to the extent that they feel more encouraged to search outside of known products that they like, with the effect that diversity also increases at the aggregate level. Indeed, as can be shown formally, if the consumer population is characterized by taste heterogeneity, a recommender system that provides personalized recommendations may lead to a 'thicker' tail in the aggregate, meaning that less-popular products receive a larger share of sales after the introduction of a recommender system.[24] A likely outcome, then, is that more niche products will be put on the market and that product variety in the market will, therefore, increase.

Oestreicher-Singer and Sundararajan (2012a, 2012b) shed some light on this issue.[25] They collected a large data set, starting in 2005, of more than 250,000 books from more than 1,400 categories sold on Amazon.com.

They restrict their analysis to categories with more than 100 books, leaving them with more than 200 categories. For all the books, they obtain detailed daily information, including copurchase links–that is, information on titles that other consumers bought together with the product in question (and which Amazon prominently communicates to consumers). These copurchase links exploit possible demand complementarities. Since these links arise from actual purchases and not from statements by consumers, they can be seen as providing reliable information about what other consumers like. By reporting these links, Amazon essentially provides a personalized shelf for each consumer according to what she was looking at last. This allows consumers to perform a directed search based on their starting point. Oestreicher-Singer and Sundararajan (2012b) find that if a copurchase relationship becomes visible, this leads, on average, to a three-fold increase in the influence that complementary products have on each others' demand.

The question, then, is how these copurchase links affect sales. In particular: which products make relative gains in such a recommendation network? Are these the products that already have mass appeal (because they are linked to

[24] See Hervas-Drane (2015) for a formal analysis.

[25] Other relevant empirical work has been done by Brynjolfsson, Hu and Simester (2011) and Elberse and Oberholzer-Gee (2007). Brynjolfsson, Hu and Simester (2011) compare online and offline retailing and find that online sales are more dispersed. While compatible with the hypothesis that recommender networks lead to more-dispersed sales, other explanations can be given. Elberse and Oberholzer-Gee (2007), comparing DVD sales in 2005 to those in 2000, find that the tail had got longer in 2005. However, they also find that a few blockbusters enjoy even more sales; this is like a superstar effect. Again, the role of recommender systems is not explicit.

other products) or, rather, niche products? To answer this question, one must measure the strength of the links that point to a particular product. For this, it is important to count the number of links pointing to a product and to know the popularity of the products from which a link originates. Hence, a web page receives a high ranking if the web pages of many other products point to it or if highly ranked pages point to it. This is measured by a weighted page rank based on Google's initial algorithm. Oestreicher-Singer and Sundararajan (2012a) construct the Gini coefficient for each product category as a measure of demand diversity within a category. They regress this measure of demand diversity on the page rank (averaged within a category), together with a number of other variables. In their 30-day sample, they find that categories with a higher page rank are associated with a significantly lower Gini coefficient. This means that in a product category in which, on average, recommendations play an important role, niche products within this category do relatively better in terms of sales, whereas popular products perform relatively worse than in a product category where this is not the case. This is seen as evidence in support of the theory of the long tail.[26]

The finding that a recommender system favors products in the long tail suggests that such a system may encourage participation on the seller side, as it becomes more attractive for niche players to become active. Since an increase in the number of buyers improves the granularity of the recommender system, a platform with a well-designed recommender system features positive cross-group external effects from buyers to marginal sellers.

Recommender systems may use information that is different from the actual purchases, but may also use hints of purchase intentions. For instance, Amazon can recommend products based on clicking behavior. If many people who looked at one product also took a close look at another product, this may suggest that the two products are closely related (as substitutes or complements) and that potential buyers benefit from cross-recommendations. We note that recommender systems may also have a future in physical retailing, provided that shoppers use a device that can provide personalized recommendations. For instance, in-shop displays may make personalized recommendations based on a shopper's history and the histories of fellow shoppers.

3. Search Engine Bias and Quality Degradation

As in the design of review and rating systems, platforms may have incentives that are not aligned with those of buyers. In particular, a profit-

[26] To take into account possible unobserved heterogeneity in the data, Oestreicher-Singer and Sundararajan (2012a) also construct a panel data set. The estimation results are confirmed with panel data techniques.

maximizing platform may have an incentive to distort the recommender system or make it less informative. The theoretical literature has uncovered several reasons that platforms operating as search engines may have an incentive to bias their search results. First, a platform may favor search results from which it can extract larger profits. Second, partial integration of the platform with some sellers or content providers may reinforce the previous motivation. Finally, a platform may discourage search so as to reduce competition among sellers. We examine these three motivations, in turn, and comment on empirical results when available.

— Search bias to favor more-profitable sellers

A platform may bias the order of recommendations if different offers lead to different commissions or to different purchase probabilities. Regarding the former, such higher margins occur if the platform has a specific partner program for which it charges higher commissions. Regarding the latter, if an offer is available on different distribution channels and some buyers multihome, these multihoming buyers are likely to purchase elsewhere if offers on alternative distribution channels are available at a lower price. Therefore, a profit-maximizing platform would place offers that were cheaper elsewhere in a lower position than if such lower-priced alternatives were not available.[27]

Given such motivations, it is interesting to ask whether platforms list search results in the best interest of consumers. Hunold, Kesler, and Laitenberger (2017) empirically investigate this issue in the context of hotel booking sites. Booking and Expedia use a default to place their recommendations–Expedia calls this list "Recommended" and Booking "Top Picks." These platforms do not provide clear information on how they construct the lists; this is in contrast to other listings that a user can obtained and that are based on price or reviewer ratings. Thus, platforms maintain discretion over how they order the available offers in the list. The authors use data from July 2016 to January 2017 from Booking, Expedia, and the meta-search site Kayak for hotels in 250 cities (most of them within Europe), featuring more than 18,000 hotels. They find that for a given price on a hotel booking platform, a lower price on the other platform or on the hotel's website leads to a worse position on the list. This suggests that hotel booking platforms bias their recommendations.

The interaction between organic and sponsored links can provide another reason that search engines opt to bias their search results–this insight is relevant

[27] If the platform is allowed to impose a most-favored nation (MFN) clause that does not allow sellers to offer lower prices elsewhere, it no longer has the incentive to bias search results in that way. However, such MFN clauses have been declared illegal in several jurisdictions on competition grounds.

not only for general search engines, but also platforms such as Booking, which offers advertising opportunities in addition to providing organic search results.[28] As Xu, Chen, and Whinston (2012), Taylor (2013), and White (2013) point out, organic links give producers a free substitute to sponsored links on the search engine. Hence, if the search engine provides high quality in its organic links, it cannibalizes its revenue from sponsored links (if it is not able to fully recoup them through higher charges on its sponsored links). At the same time, providing better (*i.e.,* more reliable) organic search results makes the search engine more attractive. If consumers have search costs, a more attractive search engine obtains a larger demand. However, if the latter effect is (partially) dominated by self-cannibalization, a search engine optimally distorts its organic search results.

Finding 11. Profit-maximizing platforms may degrade the quality of their recommender systems or provide biased recommendations. This tends to reduce the size of within-group external effects among buyers.

— Search bias due to partial integration

A misalignment of buyer and platform incentives may also be the result of partial vertical integration. In particular, this may be alleged to give rise to or exacerbate search engine bias–an issue that received prominence in the Google Shopping case in the European Union. Does partial vertical integration lead to additional worries about *search engine bias,* or can integration possibly reduce search engine bias? In what follows, we present the models of de Cornière and Taylor (2014) and Burguet, Caminal, and Ellman (2015) to systematically analyze the costs and benefits of search engine integration.

De Cornière and Taylor (2014) analyze a market with a monopoly search engine, two websites, sellers and users. The websites offer horizontally differentiated content. This is formalized by the Hotelling line, with platform 1 located at point 0 and platform 2 at point 1, and users uniformly distributed on the unit interval. Prior to search, users are not aware of their preferred content. This implies that without searching, a user cannot identify which website has the content that interests her the most. A user incurs a user-specific search cost when engaging in search on the search engine (specifically, the search cost is drawn from some cumulative distribution function).

Websites and the search engine obtain revenues exclusively from advertising posted by sellers, which users are assumed to dislike. The search engine works

[28] Our discussion of search engine bias closely follows the exposition in Peitz and Reisinger (2016).

as follows: if a user decides to use the search engine, she enters a query. The search engine then directs the user to one of the websites. The search engine's decision rule is a threshold rule such that all users to the left of the threshold are directed to platform 1 and those to the right are directed to platform 2. A key assumption is that ads on the search engine and those on the media platforms are imperfect substitutes. That is, the marginal value of an ad on one outlet decreases as the number of advertisements on the other outlet increases. This implies that the advertising revenue generated by a website falls if the amount of advertising on the search engine rises (which is treated as exogenous).

The timing of the game is as follows. First, websites choose their advertising levels and the search engine chooses the threshold. Second, the advertising market clears. Third, users decide whether or not to rely on the search engine. Finally, those users who rely on the search engine type in a query and visit the website suggested by the search engine. When deciding whether or not to rely on the search engine, a user knows the threshold and has an expectation about the websites' advertising levels. The search engine is said to be biased if its chosen threshold differs from the one that maximizes the expected user utility (and, thus, the users participation rate).

The search engine faces the following trade-off. On the one hand, it is interested in high user participation. Other things equal, a larger number of search engine users leads to higher profits because advertisers are willing to pay more to the search engine. Therefore, the search engine cares about relevance to users. In addition, since users dislike advertising, they prefer to be directed to a site that shows few ads. These considerations align the incentives of the search engine with those of users. On the other hand, the search engine obtains profits from advertisers and, thus, aims to maintain a high price for its own links. Therefore, if ads on website i are particularly good substitutes for ads on the search engine, the search engine prefers to bias results against this website.

De Cornière and Taylor (2014) then analyze the effects of integration of the search engine with one of the websites –say, website 1. Suppose that there is partial integration without control of ad levels– that is, website 1 shares a fraction ρ_1 of its profit with the search engine but retains full control with respect to its ad level (this corresponds to partial ownership, but no control rights for the search engine). Then, the search engine has an incentive to bias its result in favor of website 1 because it benefits directly from this website's revenues. However, it also benefits more from higher user participation, implying that the search engine wants to implement higher quality (*i.e.,* less-biased results). Because of these two potentially countervailing forces, partial integration can increase or decrease the level of bias. In particular, if the search engine were biased to the detriment of website 1 without integration, partial integration

might mitigate this bias. Even if the search engine is biased in favor of media outlet 1 without integration, partial integration can lead to a reduction in the bias. If the websites are symmetric, partial (or full) integration always leads to an increase in bias. However, users may be better off because of lower ad levels.

Burguet, Caminal, and Ellman (2015) propose a different setup to analyze the problem of search engine bias and integration. They do not account for ad nuisance but explicitly model consumer search for sellers' products. User i is interested in the content of one of the N websites only –this website is denoted by $n(i)$– while any other content generates a net utility of zero. Each website's content interests the same fraction of users, $1/N$.

Users do not know which website matches their interests and need the help of a search engine. Suppose that the search engine can perfectly identify the relevant website $n(i)$ once a user i has typed in the search query. When using the search engine, a user incurs a search cost.[29] The search engine displays a link to a website after a user has typed in the query. The search engine chooses the probability that the link leads to the content matching the user's interest. Since the links to websites are non-paid, this corresponds to organic search.

The search engine also features sponsored search in which it advertises the sellers products. This is the source of profits for the search engine and websites. Sellers belong to one of J different product categories, indexed by j. User i values only one category $j(i)$. Each category's products interest the same fraction of users, $1/J$. There are two sellers in each category. Seller 1 provides the best match to a user, leading to a net utility of v_1. Producer 2 provides a worse match such that $0<v_2<v_1$. The sellers' margins are m_1 and m_2. Users' and sellers' interests are assumed to be misaligned, and, thus, $m_2>m_1$. In addition, it is assumed that buyer preferences dominate for the welfare ranking –*i.e.,* $v_1+m_1>v_2+m_2$. The monopoly search engine provides a single link after a user has typed in a query for product search in a particular category.[30] Then, the search engine sets a pay-per-click price. The search engine chooses to display the link of producer 1 with some probability and the link of producer 2 with the remaining probability.[31]

[29] The search cost is heterogeneous across consumers and drawn from some cumulative distribution function.

[30] Both models described here (Burguet, Caminal and Elllman, 2014, and de Cornière and Taylor, 2014) assume that users visit only a single website after typing in a query. However, in reality users may click on multiple search results (in sequential order). They can be expected to broadly follow the respective ranking of the results. In such a situation, advertisers exert negative externalities on each other when bidding for more prominent placement. Athey and Ellison (2011) and Kempe and Mahdian (2008) study the question of how the optimal selling mechanism of the search engine takes these externalities into account.

[31] This is a simplified version of the model of Burguet, Caminal, and Ellman (2015), which is developed in Peitz and Reisinger (2016).

Absent vertical integration, search results are distorted because websites compete for advertisers. As Burguet, Caminal, and Ellman (2015) show, generically, the search engine will distort, at most, one type of search –product search or content search– setting the other at the optimal value. If the search engine was integrated with all websites, it would internalize the externality exerted by one websites on others and, as a result, improve its reliability. This is an unambiguously positive effect. However, in case the search engine is integrated only with a fraction of the websites, it has an incentive to divert search from non-affiliated websites to affiliated ones. Here, partial integration may lead to a lower consumer surplus compared to no integration.

The findings from the theoretical literature suggest that search engine bias may arise due to (partial) integration. However, partial integration sometimes is a remedy for search engine bias prior to integration, and, in any case, its consumer-welfare implications are ambiguous. So, to ascertain whether recommender systems work better or worse under (partial) integration, a detailed understanding of the specific case is needed. What is clear is that when (partial) integration reduces bias and increases buyer participation, integration tends to improve the recommender system.

Finding 12. Partial integration of a platform with sellers or content providers may increase or decrease the bias of its recommender system. Even if partial integration increases bias, it may increase buyer participation and buyer surplus.

— Search discouragement to reduce sellers' competition

Finally, a platform may want to make its recommender system less informative so as to discourage search. Chen and He (2011) and Eliaz and Spiegler (2011) provide a reason that a search engine may bias its recommendations or search results if it takes a cut from the transaction between buyer and seller–this is a situation with sponsored links. In this case, it is in the search engine's best interest for sellers' revenues from sponsored links to be high. Because revenues increase if product market competition between sellers becomes softer, the search engine may distort search results so as to relax product market competition. As formalized in Chen and He (2011) and Eliaz and Spiegler (2011), a monopoly search engine has an incentive to decrease the relevance of its search results, thereby discouraging users from searching extensively. This quality degradation leads to less competition between sellers and, thus, to higher seller revenues, which can be partly extracted by the search engine.

IV. CONCLUSION

It is our contention that one cannot understand the functioning of prominent digital platforms such as Airbnb, Amazon, Booking, Expedia, Ebay, Google Shopping and Uber without taking proper account of their rating and recommender systems.

Such systems are crucial for the performance of digital platforms for the following simple reason: potential buyers incur an opportunity cost in evaluating how products and services fare in terms of quality and how they fit their tastes; thus, they appreciate ratings, reviews and recommendations because knowing what other buyers did in the past helps them to make better-informed decisions. Rating and reviews are particularly useful for product characteristics that everyone appreciates (in terms of value for money–these characteristics may be observable prior to purchase or only after purchase, possibly only by a fraction of buyers). In the presence of taste heterogeneity, buyers benefit from personalized recommendations, which help them find their way in selecting products.

When two-sidedness is an essential feature of a digital platform, users are often keen to infer information about the reliability of the counterparties to the transactions that they may conduct on the platform. Here, rating systems can possibly steer buyers away from low-quality sellers and can discourage sellers from misbehaving. Conversely, thanks to rating systems, sellers can stay clear of problematic buyers, and buyers may have a stronger incentive to behave properly.

In this chapter, we have analyzed the economic roles that rating and recommender systems play. In particular, we have shed light on how the effectiveness of these systems depends on the joint actions of their users and designers: not only can buyers and sellers take actions that damage the functioning of rating systems, but for-profit platforms also may have an incentive to manipulate their rating and recommender systems. Finally, throughout our analysis, we have argued that rating and recommender systems are the source of positive within-group and cross-group external effects. They are, thus, in many cases, a key driver allowing a platform to attract many buyers (and, if applicable, sellers), which is an undeniable source of competitive advantage in markets with competing platforms.

BIBLIOGRAPHY

ALBANO, G., and A. LIZZERI (2001), "Strategic certification and the provision of quality," *International Economic Review,* 42: 267-283.

Anderson, C. (2006), *The Long Tail: Why the Future of Business is Selling Less of More,* Hyperion Press, New York.

Aral, S. (2014), "The problem with online ratings," *MIT Sloan Management Review,* 55: 45-52.

Athey, S., and G. Ellison (2011), "Position auctions with consumer search," *Quarterly Journal of Economics,* 126: 1213-1270.

Bajari, P., and A. Hortaçsu (2004), "Economic insights from internet auctions," *Journal of Economic Literature,* 42: 457-486.

Banerjee, A. V. (1992), "A simple model of herd behavior," *Quarterly Journal of Economics,* 107: 797-817.

Belleflamme, P., and M. Peitz (2015), *Industrial Organization: Markets and Strategies,* 2nd edition, Cambridge University Press, Cambridge.

— (2018a), *The Economics of Platforms,* book manuscript, work in progress.

— (2018b), "Platforms and network effects," in L. Corchon and M. A. Marini (eds.), *Handbook of Game Theory and Industrial Organization,* vol. II, Edward Elgar Publisher.

Bolton, G.; Greiner, B., and A. Ockenfels (2013), "Engineering trust: Reciprocity in the production of reputation information," *Management Science,* 59(2): 265-285.

Bouvard, M., and R. Levy (2016), "Two-sided reputation in certification markets," *Management Science,* forthcoming.

Brynjolfsson, E.; Hu, Y., and D. Simester (2011), "Goodbye Pareto Principle, hello long tail: The effect of search cost on the concentration of product sales," *Management Science,* 57: 1373-1386.

Burguet, R.; Caminal, R., and M. Ellman (2015), "In Google we trust?," *International Journal of Industrial Organization,* 39: 44-55.

Cabral, L., and A. Hortaçsu (2010), "The dynamics of seller reputation: Evidence from Ebay," *Journal of Industrial Economics,* 58: 54-78.

Cai, H.; Jin, G. Z.; Liu, C., and L.-A. Zhou (2014), "Seller reputation: From word-of-mouth to centralized feedback," *International Journal of Industrial Organization,* 34: 51-65.

Chen, Y., and C. He (2011), "Paid placement: Advertising and search on the internet," *Economic Journal,* 121: F309-F328.

Chevalier, J. A., and D. Mayzlin (2006), "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research,* 43: 345-354.

Cornière, A. de, and G. Taylor (2014), "Integration and search engine bias," *Rand Journal of Economics,* 45: 576-597.

Elberse, A., and F. Oberholzer-Gee (2007), Superstars and underdogs: An examination of the long tail phenomenon in video sales, *MSI Reports: Working Paper Series,* 4: 49–72.

Eliaz, K., and R. Spiegler (2011), "A simple model of search engine pricing," *Economic Journal,* 121: F329-F339.

Fleder, D., and K. Hosanagar (2009), "Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity," *Management Science,* 55: 697-712.

Fradkin, A.; Grewal, E., and D. Holtz (2017), *The determinants of online review informativeness: Evidence from field experiments on Airbnb,* unpublished manuscript.

Hervas-Drane, A. (2015), "Recommended for you: The effect of word of mouth on sales concentration," *International Journal of Research in Marketing,* 32: 207-218.

Hui, X.-A.; Saeedi, M.; Sundaresan, N., and Z. Shen (2016), "Reputation and regulations: Evidence from Ebay," *Management Science,* 62: 3604-3616.

Hunold, M.; Kesler, R., and U. Laitenberger (2017), *Hotel rankings of online travel agents, channel pricing and consumer protection,* unpublished manuscript.

Jin, G. Z., and A. Kato (2006), "Price, quality, and reputation: Evidence from an online field experiment," *Rand Journal of Economics,* 37: 983-1005.

Kempe, D., and M. Mahdian (2008), "A cascade model for advertising in sponsored search," Proceedings of the *4th International Workshop on Internet and Network Economics (WINE).*

Klein, T.; Lambertz, C., and K. Stahl (2016), "Adverse selection and moral hazard in anonymous markets," *Journal of Political Economy,* 124: 1677-1713.

Li, L. I.; Tadelis, S., and X. Zhou (2016), Buying reputation as a signal: Evidence from online marketplace, *NBER Working Paper,* No. 22584.

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.

Lizzeri, A. (1999), "Information revelation and certification intermediaries," *Rand Journal of Economics,* 30: 214-231.

Luca, M., and G. Zervas (2016), "Fake it till you make it: Reputation, competition, and Yelp review fraud," *Management Science,* 62: 3412-3427.

Mayzlin, D.; Dover, Y., and J. Chevalier (2014), "Promotional reviews: An empirical investigation of online review manipulation," *American Economic Review,* 104: 2421-2455.

McDonald, C. G., and V. C. Slawson (2002), "Reputation in an internet auction market," *Economic Inquiry,* 40: 633-650.

Melnik, M. I., and J. Alm (2002), "Does a seller's ecommerce reputation matter? Evidence from ebay auctions," *Journal of Industrial Economics,* 50: 337-349.

Milgrom, P., and J. Roberts (1986), "Price and advertising signals of product quality," *Journal of Political Economy,* 94: 796-821.

Muchnik, L.; Anal, S., and S. J. Taylor (2013), "Social influence bias," *Science,* 341: 647-651.

Nosko, C., and S. Tadelis (2015), The limits of reputation in platform markets: An empirical analysis and field experiment, *NBER Working paper,* No. 20830.

Oestreicher-Singer, G., and A. Sundararajan (2012a), "Recommendation networks and the long tail of electronic commerce," *MIS Quarterly,* 36(1): 65-83.

— (2012b), "The visible hand? Demand effects of recommendation networks in electronic markets," *Management Science,* 58: 1963-1981.

Ott, M.; Cardie, C., and J. Hancock (2012), "Estimating the prevalence of deception in online review communities," Proceedings of the *21st international conference on World Wide Web*: 201-210.

Peitz, M., and M. Reisinger (2016), "Media economics of the internet," in S. Anderson, D. Stromberg and J. Waldfogel (eds.), *Handbook of Media Economics,* vol. 1A, North Holland (2016): 445-530.

Resnick, P.; Zeckhauser, R.; Swanson, J., and K. Lockwood (2006), "The value of reputation on Ebay: A controlled experiment," *Experimental Economics,* 9: 79-101.

Tadelis, S. (2016), "Reputation and feedback systems in online platform markets," *Annual Review of Economics,* 8: 321-340.

Taylor, G. (2013), "Search quality and revenue cannibalisation by competing search engines," *Journal of Economics and Management Strategy,* 22: 445-467.

Tucker, C., and J. Zhang (2011), "How does popularity information affect choices? Theory and a field experiment," *Management Science,* 57: 828-842.

Vana, P., and A. Lambrecht (2018), *Online reviews: Star ratings, position effects and purchase likelihood,* unpublished working paper.

White, A. (2013), "Search engines: Left side quality versus right side profits," *International Journal of Industrial Organization,* 31: 690-701.

Xu, H.; Liu, D.; Wang, H., and A. Stavrou (2015), "E-commerce reputation manipulation: The emergence of reputation-escalation-as-a-service," Proceedings of *24th World Wide Web Conference (WWW 2015):* 1296-1306.

Xu, L.; Chen, J., and A. Winston (2012), "Effects of the presence of organic listing in search advertising," *Information System Research,* 23: 1284-1302.

Zervas, G.; Proserpio, D., and B. John (2015), A first look at online reputation on Airbnb, where every stay is above average, *Working paper,* Boston University.

PART II

Pricing Mechanisms and Search

PERSONALIZED PRICES IN THE DIGITAL ECONOMY[1]

Juan-José GANUZA
Gerard LLOBET

Abstract

Consumer information is becoming an increasingly important asset in the digital economy, allowing firms to offer targeted prices to consumers. This paper tries to shed some light on the economic trade-offs that arise when such information is obtained. We study the interplay between firms that use it for pricing purposes and consumers that want to prevent it from spreading out if they anticipate that it will be used to offer personalized and potentially higher prices. Finally, we study the emergence of data brokers, new platforms that gather and organize consumer information to sell to final market producers.

Key words: Consumer information, digital economy, targeted prices, digital platforms.

JEL classification: D82, L81.

[1] This paper benefited from comments by Paul Belleflamme.

1. INTRODUCTION

A tourist visiting the bazaar of an exotic town is offered products with prices that might not only be the result of a more or less tedious bargaining process, but they might also be a reflection of his/her own characteristics. Tourists with a different country of origin, age, or dress code are likely receive a different price for the same item in the same shop.

This heterogeneity in prices for the same product has been a prevalent feature throughout history.[2] The posted prices that we are familiar with and that do not distinguish among consumers are quite a historic anomaly. In towns or situations in which there was little competition, shop owners used to charge different prices to different consumers based on the information that sellers had on their characteristics learned, for example, from previous interactions. As Gordon (2016) explains, posted prices emerge as a result of the challenges that the development of the department store engendered. This modern retail model, which developed during the first half of the twenty century, along with the increasing urbanization, allowed shop owners to benefit from scale economies in their operations. More consumers could be reached and a wider variety of products could be offered while, at the same time, costs could be severely reduced. This model, however, had some drawbacks. First, markets became less local. Consumers were buying from different stores, limiting the information that sellers could gather, for example, out of previous purchases. Second, department stores required many workers to attend the growing number of customers. These workers did not have neither the information nor the experience to set prices to each consumer based on his/her characteristics, in the way that the traditional store owner used to do. The increasing lack of information stemming from the anonymity that department stores allowed, together with the need to set simple pricing rules to their workers, gave prevalence to the posted prices. Under this system, transactions were faster and workers had less discretion and required less supervision.[3] These cost savings more than compensated for the losses from not being able to discriminate prices.[4]

The internet has transformed these industries again. Our browser's cookies convey information about our preferences that data brokers sell to retailers

[2] Prices of big ticket items like cars or houses are also set as a result of negotiation, which leads to different consumers paying different prices.

[3] A commitment to posted prices could also have strategic advantages if it allowed firms, for example, to sustain uniform prices across markets. As Dobson and Waterson (2015) show, this uniformity could reduce competition, particularly in large and more profitable markets.

[4] This new paradigm also forced sellers to develop new ways to charge different prices to consumers depending on their willingness to pay. In this case, however, all consumers had to be offered the same options from which they would choose differently according to their preferences. Quantity discounts are the classic example of this kind of indirect price discrimination or menu pricing.

who complement it with their own data arising from their previous purchase history. At the same time, growing computing power and the application of big data has made easier to ascertain this information to elicit individual consumer valuations that can be used to offer personalized prices. In some sense, the digital economy allows sellers to engage globally in the kind of price discrimination that rural retailers used to be able to carry out locally.

The use of big data techniques to identify patterns in consumer preferences can be used to improve the offers that firms make to their customers. A platform, for example, can improve the recommendations that users receive depending on how much information it can gather about their preferences. However, this information can also be used to discriminate prices. In particular, Shiller (2014) shows that a platform like Netflix could in theory improve the accuracy of its predictions about the willingness to pay of consumers by tracking their browsing behavior in a few sites like Rotten Tomatoes or Wikipedia. This kind of discrimination is already used in some contexts. In a controversial case the online travel agency Orbitz admitted to displaying more expensive hotels to Mac users, since their willigness to pay was estimated to be around 30% higher.[5]

In this paper we focus on the role that information acquisition has on the prices that consumers receive. However the previous example shows, when firms act upon the consumer information they possess, many ethical implications beyond prices and consumer surplus arise. This is particularly relevant since some papers suggest that the vast amount of information about users on the internet has the potential to carve out most of their privacy. In an influential paper, Acquisti and Gross (2009) shows that, using publicly available official data combined with date of birth information from data brokers, the Social Security Number of any American resident could be guessed with a modest margin of error.[6]

The growing dimension of online price-discrimination schemes has led to an expansion of the economic literature trying to ascertain their implications for firms, consumers, and social welfare in general. In this paper we review some of its main contributions. We start with the classical literature that analyzed the effect of price discrimination in contexts in which sellers had exogenous information about consumers. We then study the incentives for firms to gather consumer information and of these consumers to trade their privacy for better or more personalized products.[7]

---

[5] See "On Orbitz, Mac Users Steered to Pricier Hotels," *Wall Street Journal* (Aug 23, 2012).

[6] We refer to Acquisti,Taylor and Wagman (2016) for a discussion of these issues.

[7] Of course, our paper is not the first one that reviews this growing literature and it complements previous surveys like Acquisti,Taylor and Wagman (2016) or Fudenberg and Villas-Boas (2012).

II. HOW FIRMS USE CUSTOMER INFORMATION?

Nowadays, it is natural to assume that firms gather information about us and learn a lot over our preferences and willingness to pay for their products. Charles Duhigg in *The New York Times* reported in 2012 a story about the use that the chain store Target made of this information. It uses as an example how the information about previous purchases is employed to predict whether a woman is pregnant and deliver coupons and offers that are directed to their new situation. As the article explained, customers are more likely to be steered towards their stores when they undergo lifetime changes and their routines change. Finding out this change at the right time is of the essence to these firms.

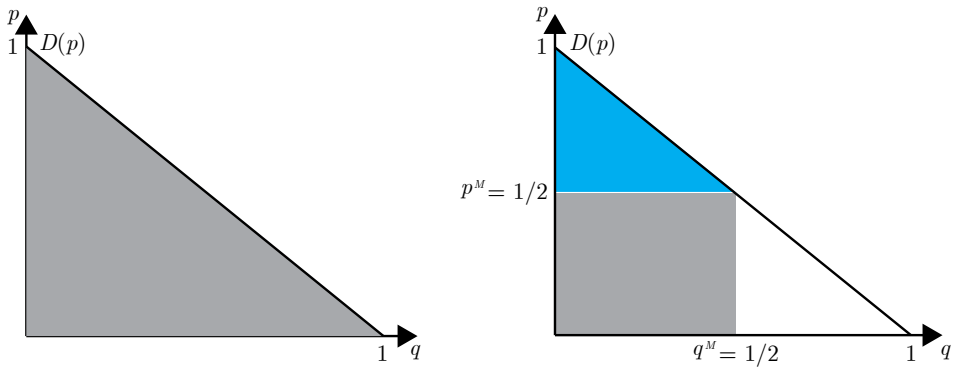
In general, a local seller like the one described in the previous section that had perfect information about each and every consumer that enters the store would be able to set a price exactly equal to his/her willingness to pay if this valuation is greater than the cost. In other words, a consumer with valuation for at most one unit of the good of v, greater than the cost of the good c, would face a price $p=v$. Importantly, all consumers that valued the good more than the cost of providing it would be served, exhausting all the benefits from trade, leading to an efficient outcome.

This efficiency result stands in contrast with the situation in which the seller has no information about consumers and is forced to charge the same price to all of them. If, for simplicity, we assume a cost $c = 0$ and consumers uniformly distributed between 0 and 1, the price that maximizes seller profits would equal 1/2. That price is lower than the valuation for the good of half of the consumers and higher for the other half. The implications for both kinds of consumers of this unique price are dramatically different. For the half of the consumers with valuation greater than 1/2 trade occurs exactly as in the situation in which the firm had all the information. The only difference, of course, is that these consumers benefit from a lower price. The other half of the consumers are harmed because the seller is not aware of their lower valuation and some gains from trade will not be realized. This efficiency loss is the well-known dead-weight loss from market power. These two results are illustrated in Figure 1.

It is clear that in this case, the more information the seller possesses the more accurate will be the purchasing decisions and the lower the dead-weight loss. Of course, the implications will be different for consumers and the firm. The former stand to lose from price discrimination. The seller extracts all the surplus from consumers with a low valuation who would otherwise not buy and, thus, they do not gain or lose from personalized prices. Higher valuation consumers, however, would surely lose since personalized prices allow the seller

FIGURE 1

CONSUMER SURPLUS (BLUE AREA) AND FIRM PROFITS (GREY AREA) UNDER PERSONALIZED PRICES (LEFT) AND UNDER A UNIQUE MONOPOLY PRICE (RIGHT)



to set higher prices for the good. For those reasons, the seller will always benefit from personalized prices.

This is a classical insight in the microeconomics literature which suggests that price discrimination has an overall positive effect on welfare and firm profits even though it harms consumers. This result also sheds some light on the incentives of consumers to protect their privacy and hinder the firm's efforts to learn their preferences and willingness to pay. We will discuss these incentives later in the paper.

In reality, a big retailer like Target has imperfect information on consumer preferences. In order to accommodate this situation we now enrich the previous basic framework and assume that the seller faces two types of consumers. Suppose that the monopolist only knows the valuation of a proportion β of these consumers and it is uninformed about the rest. This means that the firm can now post a generic price p, known to all consumers. However, because it has additional information on a subset of them, it can also offer a personalized price to this group, that we denote as $p(v)$. For this price to be relevant to informed consumers it has to be that $p(v) < p$. This kind of structure is consistent, for example, with a retailer setting a price and offering personalized discounts or coupons to consumers, attached to the usage of their loyalty card.[8]

[8] This argument assumes that there is no selection in the take up of loyalty cards. Of course, the creation of a loyalty card itself is part of an indirect price discrimination scheme. Those consumers that devote their time to fill out the form to get the card and swipe it every time they shop are also likely to be more sensitive to discounts.

At first blush, the fact that the monopolist has information over preferences to offer discounts should benefit consumers. However, this result is misleading because consumers for which the valuation is known do not benefit from this discount. In particular, those that have a valuation greater than p will never pay a personalized price greater than p and those with a lower valuation will be offered a price $p(v) = v$ which extracts all their surplus. Furthermore, the price that the monopolist will set for the general consumer increases with the proportion of consumers β for which the valuation is known. The intuition for this result is as follows. When the monopolist has no information about preferences the cost of increasing the price is the loss of low valuation consumers that will not buy. The fact that the monopolist knows the valuation of some of these consumers mitigates this cost because they will also be reached through the discounts. As a result, the optimal generic price is now $p^* = \frac{1}{2-\beta}$ which is always greater than the previous monopoly price, 1/2. In fact this case spawns the two situations discussed earlier. When $\beta = 0$ the generic price is 1/2 and when $\beta = 1$ the generic price is 1 and all consumers buy at the personalized price. For the same reasons than before, the higher is β the lower is consumer surplus out the more efficient is the final allocation.[9]

The previous discussion is consistent with physical retailers that post prices that all consumers can observe. Online retailers, however, offer a unique price to each visitor that is based on the information available, using tracking devices like browser cookies. This is the case analyzed in Belleflamme and Vergote (2016), where a monopolist charges the standard monopoly price 1/2 to consumers for which no information is available and a personalized price $p(v) = v$ for those for which the valuation is known. Notice that in this case uninformed consumers enjoy a more favorable treatment while informed consumers with a high valuation do not have access to the generic price and are worse off. Nevertheless, the general conclusion, that more information on preferences benefit the monopolist and harms consumers, goes through in this case as well.[10]

[9] The monopolist solves the following problem

$$\max_{p} (1-\beta)(1-p)p + \beta\left[(1-p)p + \int_0^p v dv\right].$$

The first term accounts for those consumers that cannot be identified and pay a price p only if their valuation is above p. The second term refers to consumers that can be identified and, in that case, if their valuation si above p they will pay the generic price p and if their valuation is lower they will receive a personalized price equal to their valuation. Equilibrium firm profits and consumer surplus are $\Pi = \frac{1}{2(2-\beta)}$ and $CS = \frac{1}{2}\left(\frac{1-\beta}{2-\beta}\right)^2$, respectively.

[10] Firm profits and consumer surplus are in this case $\Pi = \frac{1+\beta}{4}$ and $CS = \frac{1-\beta}{8}$, respectively.

The comparison of both cases, which we have exemplified as the difference between physical and online retailers, indicates that the monopolist is better off when the generic price is not available to all consumers since this allows cream skimming the market of informed consumers. By the same token, informed consumers are worse off in that case, whereas uninformed consumers face a lower generic price and they are better off. It turns out that the compounding effect is an increase in consumer surplus when the generic price is not available to all consumers. As a result, social welfare is also higher in that case. This is a surprising outcome, as it would seem that offering consumers more possibilities to choose from should be beneficial to them. As it will happen throughout this paper, however, this conclusion does not anticipate the fact that the firm will respond to its lower capability to discriminate prices by raising the generic price and, as a result, decrease consumer and social welfare.

It is important to mention that the previous discussion, focused on prices, abstracts from an important feature that the identification of consumer preferences allows. As Varian (1997) points out, firms that have this information can personalize not only the price but also the characteristics of the products that they offer, improving the match with the consumer, leading to additional gains.

III. CONSUMER INFORMATION AND COMPETITION

The previous discussion is based on the presence of a single firm, a monopolist, who could impose prices on consumers. In this setting the message is clear: price discrimination harms consumers but increases trade and social surplus. As we show next, introducing competition among sellers might reverse the result and price discrimination might benefit consumers.

We now consider a situation in which two firms compete to attract consumers in the case in which they are informed about their willingness to pay compared to when they are not. This problem is analyzed by Thisse and Vives (1988) in a context in which firms sell products that are related but different. In particular, the authors rely on the well-known linear city model. In this model, firms are located at the extremes of a line of length 1 and consumers are uniformly distributed along the line. The location, denoted as x, reflects the taste for the characteristics of the product. A consumer located at x that buys from firm 1 (located a 0) obtains a utility of $v - tx - p_1$, where v is the stand-alone value of the product, t is the disutility that buying a variety different from his/her most preferred one entails, and p_1 is the price set by the firm. If the consumer buys from firm 2, the utility becomes $v - t\left(1 - x\right) - p_2$. Firms

set prices simultaneously. For simplicity, we assume that firms do not incur in production costs and v is sufficiently large so that consumers always wish to buy one of the products.

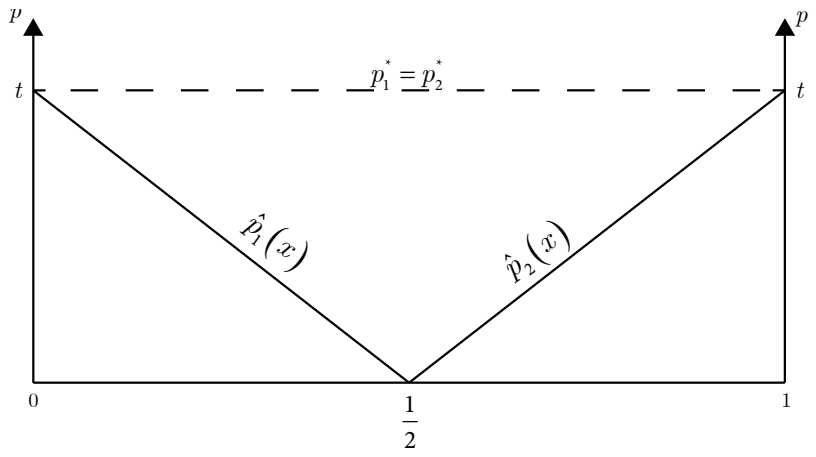
As it is standard in the literature, if firms do not have information about consumers they will choose in equilibrium a price $p_1^* = p_2^* = t$. Consumers located to the left of 1/2 will buy from firm 1 and those to the right from firm 2. The more important is the quality of the match between product attributes and consumer preferences, that is the higher is t, the larger is the market power over the agents close to a firm and the higher the price. As expected, consumers that are closer to each firm will obtain higher utility as they will incur in lower transportation costs.

If, instead, firms know perfectly consumer preferences, which in this case it is akin of knowing their location, the results change dramatically. In particular, each consumer will receive a different price depending on his/her location. In some sense, each consumer is a market. Take, for example, a consumer located at $x < 1/2$, so that he/she has a preference for firm 1. As a result, this firm can always match the deal offered by the competitor and lure the consumer at a profit. Competition among these firms will drive the price of firm 2 to 0 and firm 1 will charge a premium to the consumer equal to the savings in transportation costs from buying its product. This result implies that firm 1 will charge to those consumers to the left of 1/2 a price $\hat{p}_1(x) = t(1-2x)$ and when $x > 1/2$ the price will be zero (see Figure 2). By symmetry, firm 2 will charge a price $\hat{p}_2(x) = t(2x-1)$ to consumers with $x > 1/2$ and 0 to the rest. The comparison with the previous case indicates that consumers will pay a lower price when firms know their preferences. The reason is that in the first case, firms trade off lower sales with the possibility of charging a higher price to the more loyal customers. However, under personalized prices this trade-off does not exist. Firms can expand their market by offering lower prices to additional customers without sacrificing profits from loyal customers, which fosters competition and at the end of the day benefits consumers. Interestingly, this result also means that consumers closer to the center of the line benefit so much from competition that they obtain a higher utility that those located at the extremes even though they buy a product that is further away from their most preferred one.

The previous model constitutes an extreme illustration of the effects of competition. Taylor and Wagman (2014) show that in other models the results are more nuanced and they identify two negative effects. The first is that to the extent that competition reduces firm profits, the number of firms that enter

FIGURE 2

PRICES IN THE LINEAR CITY WHEN FIRMS DO NOT DISTINGUISH CONSUMER VALUATIONS (DASHED LINE) AND WHEN THEY CAN SET A PERSONALIZED PRICE



the market and the corresponding varieties might be reduced.[11] As a result, some consumers may end up paying more under price discrimination and buying a product that is less suitable to their preferences. The second effect appears when consumers differ in their willingness to pay for quality. Here it is easy to see that under a uniform price high valuation consumers benefit from the interest of the firm to sell also to lower valuation consumers. However, when price discrimination is possible and their higher willingness to pay is identified they will face higher prices.

The main message that emerges from this study is that the implications of knowing the willingness to pay of consumers depend on whether the competitive force dominates the greater ability to extract surplus from consumers due to the individualized prices as illustrated in the previous monopoly-setting discussion. A common feature in product differentiation models like the linear city is that the firm that sells the good that is less suitable to the consumer compensates this shortcoming by driving down the equilibrium price, benefiting that consumer. When this effect does not exist and consumer willingness to pay affects firms in a similar way the force driving competition is mitigated. To illustrate this point, Taylor and Wagman (2014) solve a model in which consumers buy multiple

[11] The authors illustrate this result with a circular city model in which the number of firms is endogenous. For the same reason as in the model above, they show that rents are lower when firms know the willingness to pay of consumers and this decrease discourages some firms from entering.

units of the same good and the private information does not affect the relative advantage of each firm but, instead, it determines the number of units that he/she is willing to buy. This model behaves similarly to the monopoly case and price discrimination benefits firms but harms consumers.[12]

IV. PRIVACY AND PRICE DISCRIMINATION

The previous discussion makes clear that consumers often have incentives to conceal information about their preferences. Tourists visiting a bazaar pretend that they are not interested in the good they are about to buy. Online users can delete the cookies in their browser if they think that being identified will lead to higher prices. However, this common practice is probably unsuccessful because consumers that do not reveal their willingness to pay are for that reason identified as high valuation ones, since had they had a lower valuation they would have been eager to reveal it.

In order to illustrate this point we return to the model in the monopoly section, as discussed in Acquisti, Taylor and Wagman (2016). Start from the situation in which the monopolist has no information on consumers and suppose that it is offering a payment r to induce consumers to disclose their willingness to pay (*e.g.,* a free service). Assume that this information can be revealed in a verifiable way. In this case, consumers face a trade-off. If they reveal their information they will receive a personalized price and they will obtain no surplus other than r. If they do not make this information available they will receive a generic price which might be lower than their willingness to pay.

This model delivers a striking outcome. In equilibrium, the monopolist offers a payment r of essentially 0 and all consumers decide to relinquish their private information. This outcome is based on the classical unraveling result of Milgrom (1981) and Grossman (1981). The intuition is as follows. Suppose that consumers expect a certain general price p. Those that have a willingness to pay below p will voluntarily disclose this information in exchange for the payment r, since they do not benefit from the purchase of the good. The monopolist infers from these decisions that those consumers for which he has no information have a valuation between p and 1. Anticipating this result, the firm will set a general price higher than p which will entice consumers with a higher valuation to also reveal their information. This unraveling process leads to a generic price

[12] Along the same line, Belleflamme *et al.* (2017) show when firms have some information consumers might be worse off. They analyze a model in which when firms do not know consumer preferences competition drives prices to cost. In that setting, consumers enjoy all the surplus. However, when firms have imperfect and different information about consumers firms obtain some market power raising the price above cost.

of 1 and provides incentives for all consumers to reveal their willingness to pay for any positive amount r. Thus, the monopolist will maximize profits by lowering this payment to essentially 0.

The previous example illustrates a situation in which consumers cannot take advantage of the control over their information. Along the same line Belleflamme and Vergote (2016) shows that the access to a technology that prevents firms from learning about consumer preferences can be detrimental to their own welfare. A discussed earlier, in that model, the monopolist tracks consumers with a certain probability and, when successful, it offers a price equal to their valuation. Otherwise, they receive a generic price. This difference spurs consumers with a high valuation to acquire a hiding technology (*e.g.,* software that eliminates cookies and erases their browsing history), since this helps them to have access to the generic price. The monopolist anticipates this behavior and, in order to discourage it, raises the generic price which makes the hiding technology useless for a larger proportion of the consumers, harming those that could never be identified. As a result, the quantity sold decreases, reducing total surplus. Consumers might be worse off overall, since those using the tracking technology may pay a lower price but a large proportion of consumers face a higher price.

The common message from these models is that, although individually consumers benefit from not disclosing information over their preferences, the impact of the strategies used to prevent the firm from learning on the final price may be self-defeating.

This negative effect extends to dynamic settings in which consumers make repeated purchasing decisions over time and firms learn from these choices about their willingness to pay. When consumers are aware of that effect they might modify strategically their purchasing decision in order to pretend that they have a lower valuation. These actions are very similar to the acquisition of a hiding technology since they will only be used by those consumers that have a high valuation. Similarly to what occurred in the models discussed earlier, when firms anticipate this behavior they will increase their future prices and harm consumers.

To illustrate this point, consider the two-period model discussed in Acquisti and Varian (2005). In that model a seller can set the price for the good in two periods. A unique consumer has a valuation constant over time that can be either high, v_H, or low, v_L, with probability π and $1 - \pi$, respectively. Because consumer valuation is constant the monopolist can use the initial period price to learn about the valuation and condition the second period price on that behavior. In order to analyze the effect of this strategy, assume first that the consumer is myopic. That is, he/she does not anticipate that the first purchasing decision can be used to extract surplus in the future once the valuation is known

and he/she is offered a personalized price. As a result, if the probability that the consumer has a high valuation is sufficiently large, $\pi > \frac{v_L}{2v_H - v_L}$, it becomes optimal to charge a high price $p_1 = v_H$ in the first period so that only a high-valuation consumer buys, uncovering this willingness to buy. The monopolist would then set a second period price equal to this consumer's valuation, so that $p_2 = v_H$ when the consumer bought in the first period and when he/she did not and, therefore the valuation was low, the price would be set to $p_2 = v_L$.

Of course, a sophisticated consumer will anticipate this ruse. If the valuation is high, buying in the first period conveys information that leads to a high price in the second period. Pretending to be a low valuation consumer implies not buying in the first period in order to obtain a low price in the second. This mechanism is a reflection of the classical "ratchet effect" described in Freixas, Guesnerie and Tirole (1985). If the firm wants to prevent this misrepresentation from happening it will have to lower the first period price. Acquisti and Varian (2005) show that the profits from doing so are lower than those from giving up on price discrimination and charging either always a price equal to v_L so that the consumer always buys or a price equal to v_H and exclude low valuation consumers.

Both scenarios are somehow extreme. When the monopolist faces a set of consumers, some of which are sophisticated and some are myopic, conditioning sales on post-purchasing decisions will typically be optimal (see also Taylor, 2004). Other reasons might also make this kind of strategy optimal. For example, products can be designed to fit certain consumer characteristics, learned from previous purchases. Consumers might then anticipate that the revelation of their valuation might entail a positive effect that could dominate the higher price that they will face.

Fudenberg and Tirole (2000) uncover another effect of conditioning on previous purchases. Contrary to the previous setup, they study a context of competition between two firms that offer differentiated products to the same consumers during two periods. These consumers have, as in the linear city, a preference for one of these products and they decide every period their purchasing decision. In second period, firms set a price for their product that they can condition on whether the agent is a returning consumer and he/she is, therefore, likely to have a high valuation for the good. As a result, in the second period the firm will charge a different price to loyal customers, that have indicated with their previous decision to have a high valuation for the good, and a low price to customers that it is aiming to poach from the other firm. The authors show that this aggressive pricing strategy in the second period leads to inefficient switching. Consumers that preferred one of the products will be attracted to the competing firm because of the good deal that they are offered. This effect will feed back into higher prices in the first period, since each

firm anticipates that attracting consumers is less profitable than in the case in which the price could not depend on previous sales. The reason is that those consumers that have a weak preference for their product will be poached in the second period. As it turns out, the total welfare implications of this strategy are negative due to the misallocation of consumers among firms and the inefficient switching.

The previous discussion abstracts from the learning that may take place over time about consumer preferences and that we highlighted in previous sections. A firm will learn from consumers that bought in the first period and this information may allow future price discrimination. This feature is important because customer poaching is based on the fact that the firm that attracted the consumers in the first period cannot retain them by discounting the price in the future when the competitor is offering a better deal. With personalized prices this discount is possible without jeopardizing the profits from more loyal customers. Choe, King and Matsushima (2016) study such a model and show that as poaching becomes less effective, competition in the first period to attract customers becomes more intense, reducing initial prices.

Poaching can also be hindered when firms use the information they gather from their consumers in order to target additional services or taylor the products to their preferences, along the lines of the discussion in Varian (1997). Zhang (2011) discusses this issue by extending the model in Fudenberg and Tirole (2000) so that firms choose not only prices but also designs (or in the model, locations). While each firm offers a unique design in the first period, it may offer a second one to new customers in the second. When segmentation of the market is posible (that is, returning customers can be prevented from accessing the design aimed at new customers) it is optimal to offer two designs. If segmentation is not feasible the offer of two designs is not an equilibrium since firms anticipate that it would lead to more competition and lower profits, for reasons that resemble those in Thisse and Vives (1988). Under segmentation each firm offers a second design that is closer to the preferences of the new customers, enticing their switching. Thus, in equilibrium customers with a weak preference for the product they bought in the first period switch to the other firm. However, contrary to Fudenberg and Tirole (2000), this switching is efficient since consumers buy a new design that is closer to their preferences than the original one.[13]

We finish this section by discussing how competition shapes the incentives for consumers to relinquish their privacy and provide information about their preferences beyond their previous purchases. Consider the case of a consumer

[13] In this discussion, for the sake of simplicity, we have set the location of the original products to the extremes of the linear city. In the paper, however, this location or design is also endogenized and the author shows that, as a result, in the first period differentiation is reduced.

who is uncertain about his/her most preferred product (*e.g.,* a new smartphone) but can devote some time online to learn about the market offerings from existing retailers. Through his/her browsing history online sellers also learn about these preferences. The more time the consumer spends online the more precise will be the estimation that both the buyer and the sellers will have, improving the match between consumer preferences and an exiting product. This kind of problem can be framed using the setup in Ganuza (2004). This paper considers a circular city in which N symmetric firms are located at the same distance from each other. Each location means a specialized design for the product. The consumer has standard preferences that linearly decay with the distance between his/her most preferred product and the location of the design chosen. The information gathered by the firm through the internet activity of the consumer translates into a public signal over the ideal product of this consumer. The timing of the game is as follows. The consumer decides first on how much information to learn which then percolates to the firms in the market. The public signal is then realized and firms make offers on designs and prices to the consumer. The purchasing decision is made and payoffs are realized.

This model shows that the provision of information implies an interesting trade-off. As in models like Thisse and Vives (1988), the firm that has a design closest to the realization of the public signal will have an advantage over the competitors and will end up selling to the consumer. The other firms will price at marginal cost and the markup of the winner will be the difference in transportation costs between the closest design and the second closest one. The more time the consumer spends online and the more precise is the information revealed, the more aligned will be the purchasing decision with the ideal one, increasing efficiency. However, more information also grants more market power to the firm that has a design closest to the realized signal, increasing firm rents. These two forces create a trade-off. The intuition is the following. Suppose that there was no information. Both the consumer and all firms would behave as if the good was homogeneous since all products would have the same probability of being the most preferred one. In that case the price would equal marginal cost but the final allocation could be very inefficient. The probability of choosing the right design would be 1/N. Under perfect information the allocation is fully efficient and the closest firm has the largest competitive advantage. If the amount of information is somewhere in the middle and the second closest firm also has a significant chance of being the most preferred one, the difference in the expected utility of these designs would also be lower. Informational rents would also decrease in that case, alongside the efficiency in the allocation.

Finally, this trade-off changes with the number of firms. More competition reduces information rents and provides more scope for the consumer to find a

better match. As a result, the consumer decides to provide more information which increases social welfare. This conclusion is in contrast with the case in which there was a monopoly seller and consumers anticipated that the information provided would be used to extract rents from them. In other words, competition mitigates the privacy concerns of consumers.

V. SELLING CONSUMER INFORMATION

Online retailers obtain information about consumers based on their previous purchases and their browsing history, as we have illustrated before. This information is a valuable asset in their efforts to understand consumer preferences and provide personalized prices. Retailers, however, are not the only firms that organize and analyze this kind of information. Data brokers, like Teradata and Acxiom, gather and sell additional information to retailers, which might be used to complement their own data in order to design more efficient personalized prices and product offerings.

Online platforms also gather information. The usage of the search engine or other services allows Google to learn about consumer habits and interests. The time that users spend in Facebook and their interaction with other users provides information to that platform about their preferences. These platforms use this information to target the advertising campaigns of their customers.[14]

Regardles of whether this information is sold or it is kept inside the firm to provide its own services, it is clear that information has become a precious commodity in the digital economy. Firms trade information directly or through services, like ads, which embed it. This is complex market in which some firms like retailers are in the demand side. Online users, that sell their personal information, usually through free or subsidized services, are clearly in the supply side. However, platforms and data brokers might be both in the demand and supply side to the extent that they acquire information used to sell services to retailers.

The existence of this market for information can, in theory, contribute to allocate data in the hands of the agent that has the highest valuation. Because the discussion over privacy is also a debate on the property rights over

[14] The Cambridge Analytica scandal, related to the US Presidential elections of 2016 and the Brexit vote, illustrates the reach of the information advantages that these platforms might grant beyond price discrimination. See https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html for more information on the development of this scandal.

consumer information, authors like Noam (1997), suggest that the standard insights related to the Coase Theorem should apply also in these markets. In other words, consumer privacy will be protected when users value it more than the firms that are willing to apply it to discriminate prices. Using this line of thought, the way that property rights are allocated among consumers and firms affects only how the surplus from their interaction is distributed but not the party that ends up controlling the information.

One caveat of the previous argument is that for the logic of the Coase Theorem to apply low transaction costs are required. Nowadays, however, consumers generally do not actively manage the information that they relinquish to all the firms with which their interact on a daily basis, since the cost of doing so is very high compared to the individual benefit each user expects to obtain. Because these costs are high and the informational requirements are often substantial, consumers become price takers of many deals and typically exchange their privacy for a zero-price service.[15]

Due to these frictions, regulators may find optimal to aggregate the preferences of consumers and act on their behalf, minimizing as a result transaction costs.[16] Shy and Stenbacka (2016) is one of the few papers that studied how privacy should be regulated. They analyze a setup in which consumers are heterogeneous in their valuation for the product and they also have a preference for one of the two firms in the market. These consumers have bought randomly from one of these firms in the past, and this firm has learned this valuation. The authors analyze three possible regulatory regimes. The first is the no privacy regime, in which there is no protection and both firms freely collect and share information about their customers. The second regime provides a weak privacy. Firms are allowed to use their information from previous sales to personalize prices but they do not have access to information on the willingness to pay of the competitor's consumers. Finally, they study a strong privacy regime in which firms cannot offer different prices to their previous consumers, but they can still poach consumers from the competitor by offering a lower price.

The paper shows that the no privacy regime, to the extent that it implies that both firms have information about all consumers, will never arise in equilibrium, since it would imply strong competition, as we have discussed in the context of earlier models. At the other extreme, strong privacy, to the extent that it does not allow firms to discriminate prices to consumers according to their valuation,

[15] This situation might change in the future as new technologies may reduce the cost of managing privacy either because they make individual data easily portable across platforms or if personal data brokers arise to act as data gatekeepers on the consumer's behalf.

[16] This is part of the new EU General Data Protection Regulation, see https://www.eugdpr.org

is detrimental to firm profits. The weak privacy regime is a compromise between price discrimination and competition that favors firm profitability.

On the contrary, welfare typically increases in the degree of privacy. This assessment is due to the fact that, as opposed to the result in Thisse and Vives (1988), consumers face switching costs when buying from a different firm. When firms can condition on the valuation of the consumer they can always inefficiency retain consumers that are mismatched. However, a strong privacy regime promotes poaching since firms cannot discriminate among their own consumers. This competition among firms generates not only lower prices but also a better match, aligning consumer welfare with total welfare.

Market mechanisms might also make difficult attain the efficient allocation of consumer information among firms even when we abstract from the transaction costs in the way consumer privacy is managed between consumers and firms. The reason for this result is, as usual, the existence of market power. Data brokers may find optimal not to sell the information to all firms that participate in the market as this reduces competition and increases the buyers' willigness to pay. This point is emphasized by Braulin and Valletti (2016). They provide a model in which a monopolist data broker must decide to how many firms, competitors in the final market, it should sell its information. In particular, they consider the case in which two firms sell a product of differentiated quality. Consumers differ in their valuation for quality, and higher quality is associated to a higher cost. Under this assumption only high valuation consumers should buy the high quality product. The authors show that if both firms have information, the allocation will be efficient. However, and along the lines of the result in Thisse and Vives (1988), they also show that in that case competition will become fierce and, therefore, the willingness to pay for the information will be low. Instead, if only one firm buys the information and can provide personalized prices, competition is weaker and more consumer surplus can be extracted. In equilibrium the data broker will sell to only one firm, meaning that producer surplus will increase at the cost of a less efficient outcome. If the high (low) quality firm gets the information, too many (too few) consumers will buy the high quality product. This intuition goes beyond the specific setup of this model, vertical differentiation, and it arises in other contexts with downstream competition where, as in the linear city (see Montes *et al.,* 2015) the more firms have access to information the lower will be their profitability.[17]

[17] The idea that exclusivity arises in equilibrium as way to maximize the willingness to pay of the downstream users is not specific to the markets for information. For example, it arises in media markets where upstream content providers prefer to sell in exclusivity to a unique media platform, as it was shown by Armstrong (1999). However, as Belleflamme *et al.* (2017) illustrate, if the data broker can fetch data of different qualities, exclusivity might not be optimal.

The previous models capture the incentives for firms to gather and sell data directly to retailers. However, online platforms have a business model that often does not involve selling the data they gather but rather to provide a channel for firms to reach consumers. These platforms acquire online space in newspapers, blogs or other pages that they use to display advertisements of their customers products. When a consumer visits one of this internet sites he/she is exposed to personalized ads. The better the information that the intermediary has on the characteristics of this user the more accurate the match and the more relevant the ads displayed will be. These intermediaries may buy the information from data brokers or may acquire it directly by providing free services to consumers, like in the case of Google and Facebook.

The base of the business model of these platforms is that they constitute a two-sided market. They match the supply of consumer attention with the demand of advertising services. The way in which platforms typically organize this market is through auctions. They first decide how much information to provide to advertisers about consumers' characteristics and firms bid in an auction to have their ads displayed. This system has nice intuitive properties. First, the advertiser that is willing to pay more to reach a specific consumer group and is more likely to win an auction is also the one that has a product with a better match with that consumer group. Second, advertisers only pay when the match is good, making more effective the investment.[18] At the same time, because consumers are offered more relevant information on products that might be closer to their needs, their disutility from ads is reduced.

De Corniere and de Nijs (2016) study this market and address two interesting questions. How much information is a platform willing to provide to advertisers? How does this information affect the prices that consumers will pay for these products? In order to answer that, they analyze a model that, in the spirit of Ganuza (2004), considers a market where N advertisers are horizontally differentiated. A monopolist platform has exclusive access to a set of consumers and allocates the ad slots displayed to them according to a second-price auction. The platform decides between two regimes. It can either provide all the information on consumer preferences or to disclose none. Advertisers first set the same price to all consumers. Every time a consumer visits a web page, the platform reveals information according to the regime specified earlier and all advertisers bid in an auction to display an ad.

Consider first the case in which no information is disclosed. In that case, all advertisers are homogeneous, they bid the same amount and the allocation

[18] Platforms typically price their services through two different schemes. They may choose a Pay-per-click (PPC) or a Pay-per-impression (also known as Cost per Mille or CPM) scheme.

is then random. As firms behave like Bertrand competitors the platform extracts all the surplus from them, which is low due to the poor match between the winner in the auction and consumer preferences. For the same reason, because firms anticipate that the utility from the match is low, they will also set a low price.

If the platform discloses all information, advertisers learn before the auction whether their match with a specific consumer is good or not. The better is the match the higher will be the bid, since the probability that the winner sells to the consumer is higher. However, because now advertisers are heterogeneous in their valuation for the consumer, competition in the auction will be weaker. Notice that compared to the previous case, the price that firms set will be higher since they anticipate that, conditioning on winning, the match with the preferences of the consumer will be better.

When deciding how much information to provide, the platform faces a trade-off. More information improves the match and the willingness to pay of the firm for the ads. However, by making advertisers heterogeneous, it decreases competition and it increases information rents. It is easy to see that this trade-off is resolved in favor of providing information when there is a sufficiently large number of advertisers. The quality of the match when information is provided increases in N, while, at the same time, the information rents decrease in N. Interestingly, this means that the price that consumers pay increases in N for two reasons. First, the disclosure regime is more likely to be implemented. Second, in this regime the larger is N the better will be the match in expected terms between the winner in the auction and the consumer. Social surplus will increase with N but due to the higher price the effect on consumer welfare is ambiguous.

The previous model treats the data obtained from all consumers in the same way. Either all information is disclosed about them or none at all. Platforms, however, might sell information in a more complex way and allow advertisers to learn more or less from consumers depending on their characteristics. Bergemann and Bonatti (2015) discuss a model that sheds light on this question. They analyze the interaction between a platform and an advertiser when consumers have a different valuation for the good that the latter sells. In particular, the advertiser is interested in obtaining more information because it allows to target its effort to those consumers with the highest valuation. The platform sets a price to identify the valuation of each consumer and the advertiser can choose the particular subset of consumers for which it wants to learn.

Interestingly, the optimal strategy of the advertiser for a given price has a intuitive pattern. The firm is more interested in learning from the consumers at

the extremes of the distribution. It is clear that knowing about those consumer with the highest valuation allows the firm to increase the advertising expenditure over them and increase profits as a result. It is more surprising, however, that the firm also prioritizes learning from those consumers with a low valuation. This is useful not only because it avoid wasting resources on them, but also because it allows to target more accurately the level of advertising to those consumers in the middle, for which no information is gathered.[19]

The advertisers' goal from learning about consumer preferences is to be able to target their campaigns accordingly. Of course, providing the right ads is not only useful because it increases the probability that a given consumer buys, but also because it reduces the nuisance cost that ads generate. A platform has to internalize this sort of cost and provide valuable content that audiences are willing to consume together with exposure to these ads. This setting is analyzed by Anderson and Gans (2011). In this paper, the authors emphasize the trade-off between the ads that the platform allows and the number of viewers of free content (*e.g.,* broadcast TV or free newspapers). In this setup, they allow consumers to be heterogeneous in two dimensions: their match with the content provided by the platform and their disutility from the ads that they receive. Those consumers for which the content is very valuable are willing to be exposed to ads even if their disutility is very large. Those consumers that obtain a low utility from the platform will be scared away by ads.

This paper characterizes the profit-maximizing advertising effort and the content provision by the platform. The work focuses on an interesting question: how this equilibrium changes as a result of the consumers' access to an ad-blocking technology. In their model consumers can access this technology at a cost. As a result, only those for which the annoyance cost is highest will use it. The platform adapts the supply of ads to this situation and the authors show that the existence of an ad-blocking technology can increase rather than decrease the amount of ads. This result, surprising at the first sight, has a clear intuition: the consumers that still face ads have lower nuisance costs. As the demand is less elastic to its exposure to adds, the platform decides to increase the advertisement level.

Of course, this change in the advertising choice has knock-on effects on the provision of content. Suppose that the platform has to decide whether to provide niche content, that is valuable to the consumers for which the match is good, or a more mass content, that is valuable for a broader audience. To the extent that ad-blockers will be more prevalent for those consumers that have a

[19] Although this is the general case, depending on the objective function, situations may arise in which the corner solutions where it is optimal to learn either from the high or the low valuation consumers only are optimal.

better match, the weight in the profits of the platform of people that have less preference for the good but have lower nuisance cost is higher. Consequently, the content that the platform will choose to provide caters a more massive audience.

The existence of ad-blocking technology has also effects on the quality of the content itself. When advertising becomes less profitable, the platform will respond by lowering its investment. This is a prediction of the model that is tested in Shiller, Waldfogel and Ryan (2017). In that paper, they show that those websites in which the use of ad-blocking technology increases the most are also those sites for which the traffic is more likely to be reduced. Despite the fact that ad-blocking makes websites with more ads more palatable to consumers, leading to an increase traffic, the fact that we observe an overall decrease suggests that the effect of a lower investment in content quality dominates.

VI. CONCLUDING REMARKS

The technologies that the internet has spawn have reduced the cost of gathering information about consumer habits and preferences. This consumer information is becoming an increasingly important asset in the digital economy. This paper tries to shed some light on the economic trade-offs that arise when such information is obtained. We study the interplay between firms that use it for pricing purposes and consumers that want to prevent it from spreading out if they anticipate that it will be used to offer personalized and potentially higher prices.

A well-established idea among practitioners and academics is that price discrimination increases social welfare because it allows consumers to buy whenever their valuation is greater than the cost of producing the good. It is also well established that consumers may benefit more from the increase in production the more competition there is among firms. Our review of the literature qualifies this point in some dimensions.

An interesting message that emerges from this review is that although fierce competition is likely to make personalized prices good for consumers, in practice there are many reasons for which this competition may not arise. First of all, information has features that resemble a natural monopoly. Firms that have more information can design better products that can be offered to consumers, which will attract more demand and produce in turn more information. Second, when these firms are intermediaries that sell advertising services they enjoy a strong competitive advantage when they have better information about consumers. Finally, and more importantly, information can

be traded and brokers will often find profitable to sell consumer data in a quasi-exclusive basis, if by doing so they can increase, for the reasons mentioned above, the value of such information.

These arguments imply that in some circumstances competition will not be effective, as the market is unlikely to spread out the information efficiently. In those circumstances, regulation may be required to guarantee that the usage of information does not constitute a significant barrier to entry that hinders competition and prevents consumers from reaping the gains from it.

Consumer effort to manage their privacy is usually not a good alternative to a regulatory response. First, there are transaction costs of managing the own information which prevent the optimal allocation of property rights to solve the problems identified above. Second, strategies aimed at preventing firms from learning about consumer preferences have often self-defeating equilibrium consequences. We have shown that firms may respond by raising prices if, by doing so, they reduce the effectiveness of this privacy strategy. Furthermore, in the context of intermediaries that obtain data in exchange for subsidized services, the value of these services may decrease if consumers make information gathering difficult.

In this paper we have tried to describe the state of the art in this area of the literature. However, there are many dimensions that we have not dealt with. Privacy raises ethical concerns beyond its market implications. The fact that two customers are treated differently by the same firm may lead to fairness concerns and spur a negative consumer reaction. This effect may limit the usage that firms make of their information or how they exploit behavioral biases to their benefit.

In addition, the literature has ignored for the most part how the increasing importance of customer information affects market structure and the optimal policy response. This is already becoming a key issue for regulatory and competition authorities in relation to dominating platforms.

Finally, an important avenue for future research is the interaction between the price discrimination policies and machine learning. These techniques will allow firms to implement more sophisticated price-discrimination schemes. They may, however, suffer a push-back from consumers if they have access to services fed by these same techniques that allow them to manage their privacy in a more effective and granular manner than the current ad-blocking technologies discussed in this paper.

BIBLIOGRAPHY

Acquisti, A., and R. Gross (2009), "Predicting Social Security Numbers from Public Data," *Proceedings of the National Academy of Sciences,* 106 (27): 10975–10980.

Acquisti, A.; Taylor, C., and L. Wagman (2016), "The Economics of Privacy," *Journal of Economic Literature,* 54 (2): 442–492.

Acquisti, A., and H. R. Varian (2005), "Conditioning Prices on Purchase History," *Marketing Science,* 24 (3): 367–381.

Anderson, S. P., and J. S. Gans (2011), "Platform Siphoning: Ad-Avoidance and Media Content," *American Economic Journal: Microeconomics,* November, 3 (4): 1–34.

Armstrong, M. (1999), "Competition in the Pay-TV Market," *Journal of the Japanese and International Economies,* December, 13 (4): 257–280.

Belleflamme, P.; Man, W.; Lamm, W., and W. Vergote (2017), "Price Discrimination under Asymmetric Profiling of Consumers," mimeo.

Belleflamme, P., and W. Vergote (2016), "Monopoly price discrimination and privacy: The hidden cost of hiding," *Economics Letters,* 149: 141–144.

Bergemann, D., and A. Bonatti (2015), "Selling Cookies," *American Economic Journal: Microeconomics,* August. 7 (3): 259–294.

Braulin, F. C., and T. Valletti (2016), "Selling customer information to competing firms," *Economics Letters,* 149 (C): 10–14.

Choe, C. H.; King, S., and N. Matsushima (2016), "Pricing with Cookies: Behavior-Based Price Discrimination and Spatial Competition," mimeo.

De Corniere, A., and R. de Nijs (2016), "Online Advertising and Privacy," *RAND Journal of Economics,* 47 (1): 48–72.

Dobson, P. W., and M. Waterson (2015), "Chain-Store Pricing Across Local Markets," *Journal of Economics and Management Strategy,* March, 14 (1): 93–119.

Duhigg, C. H. (2012), "How Companies Learn Your Secrets," *New York Times Magazine,* 19 February 2012.

Freixas, X.; Guesnerie, R., and J. Tirole (1985), "Planning under Incomplete Information and the Ratchet Effect," *Review of Economic Studies,* 52: 173–192.

Fudenberg, D., and J. Tirole (2000), "Customer Poaching and Brand Switching," *RAND Journal of Economics,* 31 (4): 634–657.

Fudenberg, D., and J. Miguel Villas-Boas (2012), "Price Dicsrimination in the digital Economy," in M. Peitz and J. Waldfogel (eds.), *The Oxford Handbook of the Digital Economy.*

Ganuza, J.-J. (2004), "Ignorance Promotes Competition: An Auction Model of Endogenous Private Valuations," *RAND Journal of Economics,* 35 (3): 583–598.

Gordon, R. J. (2016), *The Rise and Fall of American Growth,* Princeton University Press.

Grossman, S. J. (1981), "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics,* 24 (3): 461–483.

Milgrom, P. R. (1981), "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics,* 12 (2): 380–391.

Montes, R.; Sand-Zantman, W., and T. M. Valletti (2015), "The value of personal information in online markets with endogenous privacy," *TSE Working Papers,* 15: 583, Toulouse School of Economics (TSE), May.

Noam, E. (1997), "Privacy and Self-Regulation: Markets for Electronic Privacy," in *Privacy and Self-Regulation in the Information Age,* Washington, DC: US Department of Commerce, National Telecommunications and Information Administration.

Shiller, B. R. (2014), "First-Degree Price Discrimination Using Big Data," mimeo.

Shiller, B.; Waldfogel, J., and J. Ryan (2017), "Will Ad Blocking Break the Internet?," *NBER Working Papers,* 23058, National Bureau of Economic Research, Inc, January.

Shy, O., and R. Stenbacka (2016), "Customer Privacy and Competition," *Journal of Economics and Management Strategy ,* Autumn, 25 (3): 539–562.

Taylor, C. R. (2004), "Consumer Privacy and the Market for Customer Information," *RAND Journal of Economics,* 35 (4): 631–650.

Taylor, C. R., and L. Wagman (2014), "Consumer privacy in oligopolistic markets: Winners, losers, and welfare," *International Journal of Industrial Organization,* 34: 80–84.

Thisse, J.-F., and X. Vives (1988), "On The Strategic Choice of Spatial Price Policy," *The American Economic Review,* 78 (1): 122–137.

Varian, H. R. (1997), "Economics Aspects of Personal Privacy," in *Privacy and Self- Regulation in the Information Age,* Washington, DC: US Department of Commerce, National Telecommunications and Information Administration.

Zhang, J. (2011), "The Perils of Behavior-Based Personalization," *Marketing Science,* 30 (1): 170–186.

RECENT DEVELOPMENTS IN ONLINE AD AUCTIONS

Francesco DECAROLIS

Maris GOLDMANIS

Antonio PENTA

Abstract

Online advertising has been growing rapidly during the last two decades and its overall value by now exceeds that of traditional media advertisement in the US. Among the many factors behind this trend, the capacity of auction mechanisms to effectively price what advertisers' value has played a key role in shaping the behavior of the most prominent search engine and social media companies. This essay reviews how the leading auction mechanisms for online ad sales evolved over time, illustrates how these changes can be understood through the lenses of economic theory and applies the same tools to discuss some potential future developments in online ad auctions.

Key words: On line Ad Auctions, media.

JEL classification: D44, L82, L86.

I. INTRODUCTION

Online advertising is the main source of revenues for important firms such as Google, Facebook, Twitter, etc., and it represents one of the largest and fastest growing industries: in 2016, the value of online advertising (mobile and desktop) in the US alone amounted to 70 billion dollars, with an annual growth of 18%, relative to total media advertising of 179 billion dollars and an annual growth of 6.6%.[1] The vast majority of online ads are sold through auctions, in which bidders compete for the adjudication of one of a given number of 'slots' available in various online venues, such as search engine result pages, social networks feeds, magazines' webpages, and so on. Online ad auctions therefore really are the core business for one of the most important sectors of today's economy, and for many of the major and most innovative firms in what used to be called the 'new economy'.

Over the twenty years since its inception, the online ad auctions market has witnessed profound changes in its underlying auction mechanisms, the key players in the industry, and more broadly in the industrial relations. But despite representing one of the oldest and largest sectors in the high-tech industry, this market seems far from having reached a stage of maturity: this market remains very innovative, and as we will document below, it is currently undergoing important transformations, which we think are doomed to alter this important industry in a fundamental way. A good understanding of the key elements of this market, its history, and of the current forces at play, is thus crucial to understand the possible future developments of an industry in which some of the most important players of today's economy operate.

In the following, we provide a historical account of development of this market. We focus on the implications that this evolution has had for the underlying auction mechanisms adopted by the industry, and how it can be understood as a response to the changing economic environment.

II. ONLINE AD AUCTIONS: THE BASIC PROBLEM

That auctions –an economic idea which dates back at least to ancient Babylon (cf. Herodotus)– really are the core business of high-tech and super innovative firms such as Google, Facebook, Twitter, etc., may strike as odd. Yet, if one looks at the balance sheets of these firms, and looks at how the majority of revenues are generated (rather than how resources are spent and

[1] Data from Magna (2017). In 2016 in the US, the main markets for oine ads were local and national TV (67 billion dollars), radio (14 billion dollars) and newspapers and magazines (20 billion dollars).

invested), one very clear picture emerges: most of the revenues of these firms are generated by auctions.[2]

The basic auction problem is very easy to describe. In its simplest form, a seller has one object to sell; a set of possible customers submit bids, and then a rule establishes who gets the object and which price to pay. There are many variations of this basic idea, which give rise to different auction formats. The simplest and most famous of these, and the most relevant to understand the evolution of the market of online advertisement, are the following:[3]

- The (sealed-bid) *first price auction,* in which bidders submit their bids simultaneously, the highest bidder wins the object and pays a price equal to his own bid;

- The (sealed-bid) *second price auction,* in which bidders submit their bids simultaneously, the highest bidder wins the object and pays a price equal to the second-highest bid.

1. Why Auctions?

The first point which is useful to understand is why using an auction in the first place. In principle, the seller could choose a price for the object and sell it to the first customer who is willing to pay that price. The problem with this is that if the seller doesn't have a clear idea of how the demand looks like (that is, the customers' willingness to pay for that object), it is difficult to choose that price optimally: if the price is set too high, no customer would buy the object, and the seller would incur an economic loss; if the price is set too low, it may be that the object is sold at less than the maximum possible amount, and hence the seller incurs an economic loss equal to the profit forgone for not selling the object at the highest willingness to pay.

In these situations, ideally the seller would like to ask customers their willingness to pay, and then set the price optimally. But unless customers are especially naive, it should be clear that they would not respond truthfully, as they have no incentive to do so: if they did, then the seller could set the price equal to the highest valuation, thereby extracting all surplus from the

[2] In 2011, for instance, Google registered $37.9 billion in global revenues, of which $36.5 billion (96%) were attributed to advertising (Google Inc., Blake, Nosko and Tadelis, 2015).

[3] There are countless variations on these basic formats, such as the descending (Dutch) auctions, various forms of all-pay auctions, etc. Milgrom (2004), Klemperer (2004), and Krishna (2010) are excellent textbooks which discuss and analyze the main existing auction formats.

consumers. Anticipating this, customers would have an incentive to under-report their willingness to pay, to ensure that if they got the object, they would at least retain some of the surplus. This in turn makes it difficult for the seller to overcome the information problem: simply 'asking' is not enough, because the seller and the customers' incentives are not aligned. But then, what?

From the viewpoint of economic theory, auctions are essentially a sophisticated way of asking customers to reveal their valuations, but in a way which takes their incentives into account. The (sealed-bid) *second-price auction* is particularly useful to illustrate this point. As we mentioned above, in this auction bidders submit their bids simultaneously, that is without knowing the bids submitted by others, and then the highest bidder wins the object and pays a price equal to the second-highest bid. Given these rules, note that a bidder's own bid does not affect how much he pays if he wins: if bidder i wins the object, he pays the second-highest bid, not his own. Hence, in determining how much to bid, it would never be optimal to bid below his own value: by increasing his bid, he would increase the probability of winning, and still pay less than his own valuation, which would lead to an increase in his expected payoff. On the other hand, bidding above one's valuation is never a good idea: by doing that, a bidder would increase the probability of winning only in the event that the second-highest bid is above his valuation, in which case if he wins he ends up paying more than his value, incurring a loss. It follows that for all bidders in this auction it is a *dominant strategy* to place a bid equal to their own valuation. The optimal bids therefore essentially reveal bidders' willingness to pay. Moreover, since in this *equilibrium* of the auction everybody bids according to their own valuation, then the rules of the auction specify that the object goes to the agent who truly has the highest willingness to pay. In this sense, the second-price auction is *incentive compatible* (bidders, acting in their self-interest, truthfully reveal their valuation) and *efficient* (the good goes to the agent who values it the most).

As we will see below, this auction format (also known as the *Vickrey auction*), has played an important role in the development of the online ad auctions market. (Section II.3.1. provides a more detailed explanation of the second-price auction, as well as a discussion of alternative auction formats and their revenue properties).

2. Online Advertisement as an Auction Problem

At its core, the problem of online advertisement is to assign a set of objects on sale (the different slots available for advertisement on a given page), to a set of potential buyers (the advertisers). The seller in this case is the owner

of the webpage on which ads will be posted. His objective is to charge the highest possible price for each slot, but being able to do that requires knowing advertisers' willingness to pay, which as we have seen may be problematic.

In the simplest case in which a single slot is on sale, then the economic problem is essentially the same described above: every time a consumer visits a webpage, the good on sale is the advertisement slot, and the possible customers are the various advertisers interested in purchasing that slot. Of course, advertisers don't care about the slot per se. The good which advertisers really are interested in buying, through the ad slot, is the consumer's attention, and then try to transform that attention of the consumer (which has been captured by the website – be it a search engine, a magazine, or the news feed of a social network) into a sale of the product he is advertising.[4] Hence, advertisers' willingness to pay in this case can be summarized in terms of two elements: one is the expected probability that the consumer's attention is transformed into a sale, call it q_i; the other is the marginal profit made on that sale, call it π_i. Bidders' valuations in this case will therefore be equal to the expected profit generated by the presence of the ad: $v_i = q_i \cdot \pi_i$. Hence, an advertiser's willingness to pay will be larger if his per-sale-profits π_i are larger. But importantly, it will also be larger if the probability of transforming the consumer's attention into a sale is larger.

2.1. Harvesting Attention: Creating Value

It is clear that, from the viewpoint of the webpage selling advertisements, it would be best to ensure that advertisers have the highest willingness to pay possible. But there isn't much that a website could do to increase π_i: advertisers' per-sale-profits depend on their costs and prices. On the contrary, there are several things a webpage could do to increase its ability to attract customers, and to increase the probability that their attention transforms into sales (that is, to increase q_i). Understanding this point is important to understand many aspects of the development of this market. We thus list some of the most important things that a webpage could do to increase its ability to create value:

- *Make the content of the page more interesting:* Clearly, if more consumers visit a given page, then the seller will have more goods to sell (more consumers' attention). Hence, the primary objective of a webpage is to attract as many visitors as possible, because it increases the total volume of 'goods' he can sell. Ultimately, it is the intrinsic quality of the website, the interest it manages to create, which determines the volume of its ads sales.

[4] Wu (2016) provides a thorough and pleasant-to-read account of the history of advertisement.

- *Capture the attention of the consumers:* Even when a webpage attracts many visitors, it is often the case that they are not particularly attentive. If consumers are visiting the webpage in a distracted way, they will not be paying much attention to the ads either. Hence, the probability that these consumers will end up purchasing the product advertised in the ad would be lower than if the website managed to keep its visitors engaged. Improving the contents of the webpage, its layout, and usability, are key elements to maintain a high level of attention from the consumers, and hence to increase the probability that visits would ultimately translate into sales. In other words, the ability to harvest the attention of the consumers is crucial to increase q_{ii} and hence to increase the value for the advertisers.

- *Targeting, i.e., matching the right consumers to the right advertisers:* A major difference between online relative to traditional media ad is the greater *targeting* potential of the former relative to the latter. Targeting refers to the possibility of tailoring the ad to (nearly) a specific consumer. This is based on the ability to know or infer consumers' characteristics from a broad set of features. These features range from basic information, such as the geographical location of the device where the ad will be shown, to possibly detailed information on demographic characteristics of the consumer, if not even its past online behavior. It is clear that the closer the content of the page is to the product sold by a particular advertiser, the higher will be its expected number of sales. For instance, holding everything else constant, the probability of sale generated by an ad for a car dealer is likely to be higher when placed on the webpage of a car magazine, than when it's placed on the webpage of a horse magazine. To a large extent, this problem is for the advertisers to solve, by targeting the right webpages which are more likely to attract the 'right' kind of consumers. But webpages and providers have an active role in this too. First, by shaping their contents and layout, webpages affect which kind of consumers they attract, and hence ultimately the advertisers they will eventually cater to. Also, websites are constantly developing techniques to provide advertisers with increasingly accurate profiles of the consumers who visit their places. By targeting a slot to particular characteristics of the consumers (geographic area, cookies, etc.), these webpages are able to generate auctions which are particularly valuable to the potential advertisers, because they offer a higher probability that the consumers' attention will ultimately generate a sale.

- *Choose the position and size of the ads adequately:* The position of the ad space is crucial to determine its effectiveness. Consumers will typically be exposed to ads while visiting web sites looking for their 'organic' (*i.e.*,

non-ad) content. Clearly, if ads are placed at the bottom of the page, after all the organic content is over, or if they are extremely small, they won't be able to capture much attention from many customers, and hence the value for the advertisers will be smaller. In contrast, a large advertisement at the top of the page, or in the middle of an article, or well-integrated in the organic content, is very likely to be noticed, and hence capture a lot of attention, and increase the value for the advertisers. Finally, the device places a key role: ads shown on mobile devices must be different from those shown on computer screens. This affects not only the size of the ad but also more fundamental aspects such as the differential effectiveness of presenting videos, pictures or text messages.

Of course, all this is easier said than done, and as usual in economics there are trade-offs. For instance, very large ads placed in the middle of a webpage's organic content may be very effective at being noticed, but they would lower the overall quality of the webpage, and hence attract less customers or be less effective at keeping them engaged. In contrast, a very clean webpage with a good content is likely to attract many consumers, but at the cost of making the ads less likely to be noticed. Similarly, adding more ads slots increases the number of goods on sale, and hence the potential revenues, but it also decreases the effectiveness of any given one of them (different slots on the same page compete for the customers' attention, decreasing each other's expected number of sales), and the effect on the total revenues may be uncertain.

All these considerations point at crucial decisions on how to structure a given webpage, how many ads to allocate and where to position them. Making the 'optimal' choice is complicated, and requires a careful understanding of the way consumers allocate their attention on different parts of a webpage, and among different webpages. This is one of the reasons why large firms such as Google, Yahoo!, Microsoft, Facebook, Telefonica, etc., are investing huge amounts of resources in maintaining active research departments filled with economists, statisticians and computer scientists, whose efforts are dedicated in a large fraction to understanding consumers' behavior on the Internet.

But whatever the choices of how many slots to put on sale, where to position them, and next to which organic content, the remaining economic problem is that of an auction: there is a given number of goods on sale (the ads slots), and a number of potential buyers (the advertisers), with valuations that are unknown to the seller (the webpage). For this reason, another crucial activity of the research groups of the most important firms in this industry is precisely to improve the auction mechanisms used to sell advertisement space. As we will see below, much of the innovation in the area of auctions in recent years has in fact come from these private research groups.

2.2. Selling Attention: Pay-Per-Click or Pay-Per-Impression?

When a single slot is on sale on a given webpage, the seller (typically the owner of the webpage, sometimes the provider) may use some variation of the basic auction formats described above: for instance, both a first-price and a second-price auction could be used. But given the particular goods on sale, further choices need to be made. For instance, besides advertising products the way that standard commercials do (that is, by presenting them in an interesting and attractive way), online ads typically provide a short message with a link to the advertisers' websites. In fact, in some cases only the second element is present: for instance, the sponsored links on the search result pages of most major search engines nowadays do not contain a classical advertisement, they only provide a link to the advertiser's website. This means that, besides choosing the auction format (*e.g.,* first- or second-price auction), the website can now choose whether to just sell the space of the ad, or the clicks. In other words, the seller can choose whether an advertiser who has occupied a particular slot should pay for just being there (*pay-per-impression*), or should pay only when a consumer clicks on its ad (*pay-per-click*).[5]

It should be clear that, conditional on a single consumer visiting the webpage (for instance, if a new auction is generated every time that a new consumer visits the page – as is for instance the case for search engines, in which every search generates a separate auction), then the advertiser's willingness to pay-per-click is higher than his willingness to pay-per-impression, since the probability that a click turns into a sale is higher than the probability that a visit, which may even overlook the ad, leads to a sale. But in some cases sales need not go through clicking on the ad, as for instance when a user sees the ad of a particular car model on a magazine's webpage, and then he decides to buy a car at a nearby car dealer, without clicking on the ad. If advertisers have correct expectations over the probabilities that visits or clicks transform into sales, there is no reason to expect systematic effects on the expected revenues one way or the other. In fact, different webpages opt for different solutions: some choose a pay-per-click scheme (that is, they essentially sell clicks), others charge on a pay-per-impression basis (that is, they sell probabilities of clicks). By and large, price-per-click schemes tend to be preferred by webpages in which ads are limited to a link, without conveying much information or rich intrinsic content (this is the case, for instance, for the sponsored links sold on search engines' result pages). Webpages which instead allow larger advertisement space, with flashy ads and a richer content, are more likely to adopt a pay-per-impression scheme.

[5] A third system sometimes used is known as "pay-per-engagement" and entails an advertiser's payment only when the consumer actively engages with the ad. We will focus on the former two systems as pay-per-engagement is less frequently used.

2.3. The Main Online Ads Auction Formats

Online ad auctions typically involve many slots on sale on the same webpage, each to be assigned to a distinct advertiser. These auction formats are thus typically 'multi-unit' auctions, which makes the problem of assigning the right slot to the right advertiser at the best price possible much more complicated than when a single good is on sale. As we will see, this problem has lead to the creation of novel auction formats.

In the general online auction problem, there will therefore be a number of slots on sale, denoted by $s = 1, ..., S$, and a number of possible buyers (the advertisers), denoted by $i = 1, ..., n$. To make the problem interesting, we will assume that there are more advertisers than available slots, and hence that $n > S$. Slots differ in terms of the number of clicks they generate: the click-through-rate (CTR) of slot s is denoted by x_s, and represents the number of clicks that an advertisement placed in a particular slot is able to generate. Slots are numbered in terms of their CTRs, with the first slot being the best, and the last slot being the worst (that is, CTRs are ordered so that $x_1 > x_2 > ... > x_S > 0$). For simplicity, it will be useful to assume that advertisers know the CTRs associated to the various slot, and hence that they share the same ranking over the slots: holding everything else constant, they all agree that the first slot is the best, with a CTR of x_1; then the second, with CTR x_2; and so on. Advertisers' valuation, which in this case represent their willingness to pay-per-click, will be denoted by v_i as above. It will be useful to label advertisers in order of their valuations, so that bidder 1 is the one with the highest willingness to pay, bidder 2 is the one with the second-highest willingness to pay, and so on (that is, $v_1 > v_2 > ... > v_n > 0$).

In the next sections, we will focus on three main auction formats, which have played an important role in the evolution of this market. In historical order of appearance in the market, these are: (i) The *Generalized First-Price Auction;* (ii) The *Generalized Second-Price Auction;* and the (iii) The *Vickrey-Clarke-Groves Auction.* The first two auction formats typically operate on a pay-per-click basis. The third instead is often used as a pay-per-impression system.[6]

In all these auctions, bidders submit bids simultaneously (that is, without knowing others' bids, as is the case in the 'sealed-bid' basic auction formats introduced earlier). Bids are ranked from the highest to the lowest, and then the highest bidder obtains the best (first) slot; the second highest bidder obtains the second slot, and so on. So, for instance, if bidder i has placed the k-highest

[6] Both GSP and VCG auctions can in principle be implemented within pay-per-click, pay-per-impression or pay-per-experience systems.

bid, then he gets the k-highest slot, and if he pays a price-per-click p_k for the second slot, then its expected payoff is equal to $x_k \cdot (v_i - p_k)$. These auction formats only differ in the price paid for each slot, in the following way:

- In the *Generalized First-Price (GFP) Auction,* the k-th highest bidder gets the k-th slot, and pays a price-per-click equal to his own bid.

- In the *Generalized Second-Price (GSP) Auction,* the k-th highest bidder gets the k-th slot, and pays a price-per-click equal to the next (the $k+1$-th) highest bid.

- In the *Vickrey-Clarke-Groves (VCG) Auction,* the k-th highest bidder gets the k-th slot, and pays a price equal to a weighted sum of all lower bids, where weight of the l-th highest bid (for $l > k$) is equal to $(x_{l-1} - x_l)$, where we set $x_k = 0$ for all $k > S$.

These auction rules are obviously more complicated than the baseline auction formats introduced at the beginning of Section II, and we will explain them in detail in the next sections. For now, we limit ourselves to noting that when there is a single slot on sale (that is, if $S = 1$), then the GFP and GSP auction coincide, respectively, with the basic first- and second-price auctions introduced earlier. In this sense they provide a generalization of those auction formats to the case of multiple goods (hence their names).

We also note that –albeit it's perhaps harder to see– in the case of a single good ($S = 1$) the VCG auction also coincides with the baseline second-price auction. In this sense, the GSP and VCG auctions provide alternative ways of extending the baseline second-price auction to the case of multiple goods.

Sections III-IV will provide a brief history of the evolution of this market and an explanation of the three main auction formats we just introduced. Readers who are interested in grasping the economics underlying these complex auction formats are encouraged to take a short detour on basic elements of auction theory, which we provide in the next subsection. The content of Section II.3., will be useful to understand our more in-depth discussions of the GFP, GSP, and VCG auctions (respectively in Sections III.1.1., III.2.1., and III.3.1.). Readers who are only interested in the historical account may skip these sections, without impairing the readability of the rest of the article. In any case, our more in-depth discussions in Sections II.3., III.1.1., III.2.1., and III.3.1., won't require any specific technical or mathematical knowledge.

3. Basic Elements of Auction Theory

We have mentioned earlier that, from the viewpoint of economic theory, auctions are essentially a sophisticated way to ask customers to reveal their valuation, but in a way which takes their incentives into account. In this Section we explain why this is the case by introducing basic elements of auction theory, focusing on the auction formats which are most relevant to understand the evolution of online advertisement auctions: namely, the *second-price* and the *first-price auction* formats.

3.1. The Second-Price Auction

As we mentioned above, in a (sealed-bid) *second-price auction* bidders submit their bids simultaneously, that is without knowing the bids submitted by others, and then the highest bidder wins the object and pays a price equal to the second-highest bid. As we will see, this auction is particularly useful to understand in what sense auctions are 'sophisticated ways to ask customers what their willingness to pay really is'.

First note that, given the rules of the auction, a bidder's own bid does not affect how much he pays if he wins: if bidder i wins the object (which means that his own bid, b_i, was the highest of all), then the price he pays is determined by the next (second-) highest bid (call it b^*). The effect of i's own bid therefore is only to determine whether or not he wins the object (he wins if $b_i > b^*$, not otherwise),[7] not how much he pays if he wins (which is b^*, if $b_i > b^*$).

Second, note that the only thing that i cares about, besides his own bid and his valuation, is the highest bid placed by his opponents (call it b^*): if his own bid is less than the highest bid among the opponents ($b_i < b^*$), then he doesn't get the object and obtains zero. If instead his bid is higher than the highest opponents' bid ($b_i > b^*$), then he wins the object and pays the highest opponents' bid, for a total surplus of $v_i - b^*$. All other bids, of the other bidders who bid less than b^*, do not affect the payoff of bidder i. Hence, it is as if a bidder is only facing *one* opponent, rather than many: the only one that matters is the highest bidder among the others.

Now, suppose that i's own valuation for the good on sale is v_i. We show next that placing a bid equal to one's own valuation in this auction is better

[7] Throughout this article, we ignore the case of ties, in which, for instance, $b_i = b^*$. For those cases, real-world auctions normally specify tie-breaking rules. These rules often assign the good with equal probability to the bidders who tie at the top, but different tie-breaking rules are also used.

than placing any other bid. We begin by first showing that bidding one's own valuation is better than bidding below it. Let $b_i^T = v_i$ denote the 'truthful bid', and $\hat{b}_i < v_i$ some candidate lower bid. Note that for all $b^* > v_i$ and $b^* < \hat{b}_i$, the two bids b_i^T and $\hat{b}_i$ result in the same utility: 0 in the first case (if $b^* > v_i$, i does not obtain the object under both $\hat{b}_i$ and b_i^T, because the highest overall bid is b^*); $v_i - b^*$ in the second case (if $b^* < \hat{b}_i$, then i wins the object whether he bids $\hat{b}_i$ or b^T, and in both cases he pays a price equal to b^*). Hence, whether b^T is overall better than $\hat{b}_i$ depends on how the two fare against opponents' bids for which b^* falls between $\hat{b}_i$ and b_i^T. For such values of b^*, bidding $\hat{b}_i$ yields a payoff of 0, because i would lose the object; bidding truthfully instead yields a payoff of $v_i - b^*$, because i wins the object and pays the next highest bid, b^*. But since, in the situation we are considering, $b^* < v_i$, this surplus $v_i - b^*$ is larger than 0. Hence, overall we found that: for any bid below one's own valuation, $\hat{b}_i < v_i$, truthful bidding is just as good if the highest opponent bid is larger than v_i or lower than $\hat{b}_i$, but for all situations in which it is in between, the truthful bid ensures a strictly higher payoff than the underbidding strategy $\hat{b}_i$.

A similar argument shows that truthful bidding is also better than bidding above one's own valuation, $\hat{b}_i > v_i$. For all cases in which the highest opponent's bid, b^*, is larger than $\hat{b}_i$ or smaller than v_i, the two strategies induce the same surplus (zero in the former case, and $v_i - b^*$ in the second), but for the intermediate cases truthful bidding does strictly better: it yields a payoff of zero (if $b^T < b^*$, i does not get the object), whereas overbidding induces a loss.

We have thus established that, for every bidder, bidding truthfully is what game theorists call a *dominant strategy:* it is optimal, no matter what the others do. So, if all bidders act in their own self-interest, their bids will be equal to their true valuations, and in this sense the second-price auctions is nothing but a sophisticated way to ask customers what is their true willingness to pay. Moreover, since in this *equilibrium* of the auction everybody bids according to their own valuation, then the rules of the auction specify that the object goes to the agent who truly has the highest willingness to pay (call it agent 1, with valuation v_1), and that he pays a price equal to the second-highest valuation (call it v_2). The resulting allocation is therefore *efficient* (the good goes to the agent who values it the most), the seller's revenue is v_2, and the winner of the auction obtains a surplus equal to $v_1 - v_2 > 0$.

We conclude the discussion of the second-price auction with one remark which will be useful to understand an important property of the auctions used to sell online advertisement space. In particular, note that the argument above implies that bidding truthfully would remain optimal even if bidders learnt others' valuations. For instance, suppose that the same auction is repeated over time, always with the same set of bidders and with the same valuations. If these

bidders bid truthfully in every period, and past bids are observed, then over time bidders would know which bids to expect from others. Yet, they would have no incentive to lower their own bids. That is because own bids do not affect one's own payment.[8]

Advanced Section: The Optimal Auction. An attentive reader may wonder whether, since the seller is not extracting the full surplus from the winner in the second-price auction, his revenues may be improved by switching to a different selling system. We have already argued that the first-best (which would be equal to charging a price of v_1) may not be achievable in this setting. One may thus wonder what the second-best looks like: if v_2 is very low, for instance, is there an auction which guarantees higher revenues than the one described above? Economic theory does provide an answer: the *optimal auction* in this case is a second-price auction with a reservation price p^*. The rules are such that if all bids are below p^*, then the object is not sold; if the highest bid is above p^* and the second is below p^*, then the winner gets the object and pays p^*; otherwise the rules are the same (effectively, it is as if p^* is the bid of the seller). It can be shown that, if the reservation price is chosen optimally (so as to trade-off the loss incurred if $v_1 < p_1^*$, so that the object is not sold, with the gain generated when $v_1 > p^* > v_2$), then the resulting *second-price auction with the 'optimal' reservation price* still provides an effective way of eliciting bidders' valuation in a way which maximizes the expected revenues of the seller. In this sense it is the *optimal auction*.[9]

3.2. The First-Price Auction

For later reference, it will be useful to discuss another common auction format, which perhaps is the most intuitive for a non-economist: the (sealed-bid) *first-price auction*. As already mentioned, in this auction bidders submit their bids simultaneously (that is, without knowing others' bids), and then the highest bidder wins the object and pays a price equal to his own bid.

[8] Clearly, in this hypothetical situation in which valuations become known, the seller would be tempted to stop running the auction and sell the good for a posted price equal to v_1, so as to extract the entire surplus. In this discussion we are assuming that the seller at this point is *committed* to using an auction. The reason is that if he were not, and bidders realized that, they would understand that their bids would reveal their valuation and might be used against them in the future. If this were the case, then bidding truthfully wouldn't be optimal anymore, and the seller would be back to square one. In a repeated environment, commitment is therefore important for the seller to solve the information problem in the first place.

[9] Since the logic of the optimal auction is one which trades-off the probability that the highest valuation is lower than the reservation price, with the probability that the reservation price falls between the highest and second-highest valuation, it is clear that the optimal auction can be determined only if the seller has well-formed beliefs on the distribution of bidders' valuations. It is also clear that it produces potential inefficiencies, as sometimes there will be no sale even in the presence of advertisers with a positive valuation (that is, when valuations are all below the reservation price).

The fact that the winner pays his own bid in this auction complicates the strategic analysis: unlike the second-price auction, now a bidder's bid determines both his probability of winning, and the price he pays if he wins. Setting a bid equal to one's own valuation is not optimal anymore: if someone else bids above, the bidder does not win the object and obtains zero; but if everyone bids lower, then the bidder wins and pays his own bid –equal to his valuation– and hence he obtains zero. Bidding one's own valuation therefore ensures that the payoff is zero, no matter what the others do. But this in turn means that any bid $b_i < v_i$ is better than bidding truthfully for bidder i: no matter how small the probability of winning might be, say $\varepsilon > 0$, in case of victory it would yield a payoff of $v_i - b_i$, and hence in expectation it is equal to $\varepsilon \cdot (v_i - b_i) > 0$, which is still larger than the payoff obtained by bidding truthfully. Hence, if bidders are rational, they would not bid truthfully in a first-price auction.

If bidders are uncertain over other bidders' valuations –say every bidder i expects other bidders' valuations, v_j, to be drawn independently from a certain distribution $F(\cdot)$– then economic theory allows to calculate the 'equilibrium' bids. For instance, if there are two bidders, and valuations are independently drawn from a uniform distribution over [0,1] (that is, for any p, the probability that i's valuation is equal to p or less is exactly equal to p), then the equilibrium bids in this auction are such that $b_i^* = \frac{v_i}{2}$ (with n bidders, the equilibrium bids would be $b_i^* = \frac{n-1}{n} v_i$). Note that, if bidders bid according to this equilibrium, it is still the case that the highest bid is placed by the highest valuation bidder, and hence the ultimate allocation is efficient: the good goes to the highest valuation bidder, just as in the second-price auction. But what about revenues? Since bids in the first price auction determine both the probability of winning and the payment itself (the first provides a reason to increase one's bid up to his own valuation; the second provides a reason to keep one's bid as low as possible), in general equilibrium bids in the first-price auction are going to be lower than in the second-price auction. In the latter auction, however, revenues are equal to the second-highest bid, whereas in the first-price auction they are equal to the highest bid of all. Hence, the overall effect on revenues is unclear. One surprising and famous result in economic theory –*the revenue equivalence* theorem, due to 2007 Nobel laureate Roger Myerson– is that the expected revenues in these two auctions are the same (Myerson's theorem in fact is much more general than that).[10]

Hence, in summary, the first- and second-price auction induce the same allocations and the same expected revenues, but strategic behavior is much

[10] While more general, this result holds under some precise condition. Being beyond the scope of this essay, we defer a discussion of such conditions to the more technical literature.

simpler in the second-price auction, and it's more 'robust' to varying bidders' information about others.

III. BRIEF HISTORY AND ECONOMIC ANALYSIS OF THE SPONSORED SEARCH AUCTIONS

In this Section we present a brief history of the evolution of the main auction formats used in the online ad market, with a particular emphasis on its most important kind: the sponsored search auctions. In the typical sponsored search auction, an advertiser with an account with a search engine provider (like Google or Microsoft-Bing) selects for each "keyword" (a single word or a phrase) the message it would like to display, the maximum price it is willing to pay (per click or impression) and the overall budget available, as well as any targeting option that might be available. Section III.1 discusses the very early days of this market, when in 1998 the search engine GoTo.com –later renamed Overture and acquired by Yahoo! in 2001– introduced the so called *Generalized First Price* (GFP) auction to sell advertisement space on its search results pages. We discuss the key economic properties of this auction format, which also provide an explanation for its eventual dismissal.

Section III.2 instead focuses on the *Generalized Second Price* (GSP) auction, which was introduced in 2002 by Google as part of its *AdWords Select* bidding platform, and which has since been adopted by all major search engines and has become the auction format of reference in this industry. We discuss the economic properties of this auction format, its advantages over the GFP format which preceded it, and the reason of its success.

Section III.3 discusses the *Vickrey-Clarke-Groves* (VCG) auction. Unlike the GFP and GSP auctions, which were developed by private firms active in this market, the VCG auction is an old auction format which had been developed by academic economists in the early '60s, to solve the general problem of achieving an efficient allocation of goods (see Vickrey, 1961). Despite being very well-known to economic theorists, and perhaps due to its fairly complex payment rule (see Section II.2.3.), this auction format remained pretty much confined to advanced economics textbooks, until Facebook decided to adopt it. As we will see, Facebook's decision was received with a certain surprise by the industry, which could not see clear reasons to favor such a complex mechanism over the simpler available alternatives. Since then, Facebook's excellent performance suggests that the VCG auction has performed very well, and there are rumours in the industry that Google is experimenting it on some of its auctions, and possibly consider a full switch.

In Section IV we will discuss other recent developments in this market, and suggest an economic explanation for the success of the VCG auction as well as some possible implications of these recent developments for the future of this important industry.

1. Pre-history: Overture and the GFP Auction

In 1998, the search engine GoTo.com revolutionized the world of online advertising by introducing auctions to sell ad space on its search results pages. This company, later renamed Overture and acquired by Yahoo! in 2001, had devised the so called *Generalized First Price* (GFP) auction, in which advertisement space was assigned to advertisers by the ranking of their bids, with each advertiser paying his own bid for each click he received. The key idea was to realize that a search engine was able to harvest a very valuable good: consumers' attention. The next step was then to turn every search on the search engine into an auction. The scheme first developed by GoTo.com-Yahoo!, and subsequently followed by all other search engines, was essentially to generate a distinct auction for every keyword searched on the search engine.

In Yahoo!'s original format, the GFP auction, slots were assigned to bidders in decreasing order of bids (the best slot to the highest-bidder, the second slot to the second-highest bidder, and so on), and every bidder paid a price-per-click equal to his own bid. Hence, suppose that the n bidders submit a profile of bids $b = (b_1, b_2, ..., b_n)$, and i's bid is the k-highest, then he obtains slot k and pays a price-per-click equal to his own bid. The resulting payoff for this bidder is therefore $x_k \cdot (v_i - b_i)$. Each advertiser is thus restricted to one bid per keyword, without the possibility of indicating a different price for different slots.[11]

This auction format was initially very successful. Yahoo!'s revenues and capitalization grew very quickly. But as Yahoo!'s auctions grew in volume, and advertisers became acquainted with their operation, this initially very successful model became problematic (see, for instance, Ottaviani, 2003). The reason is that, after an initial period in which advertisers cycled through phases of aggressive and conservative bidding, their bids eventually settled at very low levels. This meant more volatile and overall lower revenues for Yahoo!, which was therefore vulnerable to competition from other search engines which could devise better auction formats. But to understand which features of an auction would make it overcome this kind of problems, it is important to first understand

[11] An alternative that has been experimented by search engines, but without ultimately being adopted on a large scale, involved a form of "combinatorial bidding" allowing advertiser to bid either for a regular slot or for a larger slot containing not only a short text message, but also a larger picture.

why the GFP auction may have generated these phenomena of bidding cycles and implicit bid collusion. For this reason, we turn next to an economic analysis of the GFP auction.

1.1. Economic Analysis of the GFP Auction

Similar to the baseline (single-good) first-price auction, it can be shown that when advertisers are uncertain about others' valuations, there exists an equilibrium of the GFP in which slots are assigned *efficiently*. Namely, bidding strategies such that the resulting equilibrium bids $\hat{b} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_n)$ have the property that $\hat{b}_1 > \hat{b}_2 > \ldots > \hat{b}_n$, so that the highest valuation bidder (bidder 1) obtains the best slot, the second-highest valuation bidder (bidder 2) obtains the second slot, etc. This way, for all bidders who do get a slot (namely, bidders $i = 1, \ldots, S$), they each pay their own bid $\hat{b}_i$, and the resulting payoffs are $x_i \cdot (v_i - \hat{b}_i)$.

Now, suppose that –for a given keyword-auction– the set of bidders and their valuations are fairly constant over time. If this is the case, then bidders would come to expect each other's equilibrium bids to be more or less equal to $\hat{b} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_n)$. But now consider the problem of bidder S, the one obtaining the lowest slot on sale: his payoff when everybody bids in this way is $x_S \cdot (v_S - \hat{b}_n)$. Since in the GFP auction the price-per-click is equal to a bidder's own bid, this payoff is decreasing in S's own bid. Hence, ideally this bidder would like to lower his bid as much as possible, but without losing his slot. This means that he clearly cannot just set his bid to zero, or he would lose his slot. But if the profile of bids $\hat{b} = (\hat{b}_1, \hat{b}_2 \ldots, \hat{b}_n)$ is fairly stable, then this bidder knows that he would still obtain the same slot as long as he places a bid higher than the next lower bid, $\hat{b}_{S+1}$. Thus, bidder S would have an incentive to lower his own bid as long as this happens without losing his position. If nobody changes their bid in the meantime, this ideally would be all the way down to $\hat{b}_{S+1} + \varepsilon$ (where we take $\varepsilon > 0$ to be the smallest bid increment, *e.g.,* a euro cent).

It should be clear that the logic of this argument in fact applies to every bidder i: each i would obtain the i-th slot as long as $b_i > b_{i+1}$. But apart from that, one's payoff from obtaining the i-th slot is maximized if b_i is set to the lowest possible value which ensures that i obtains his 'right' position. This means that, from an initial period of bids more or less stable at $\hat{b} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_n)$, the payment structure of the GFP gives bidders strict incentives to start lowering their bids.

But now suppose that bidder i's bid has been lowered as much as possible, without conceding his slot (*e.g.,* suppose that $b_i = \hat{b}_{i+1} + \varepsilon$). At this point, bidder

$i+1$ obtains the $i+1$-th slot at a price equal to $\hat{b}_i$, paying essentially the same as what that bidder i is paying for the i-th slot, which has a higher CTR. Hence, bidder $i+1$ would have an incentive to increase his bid over b_i (say, to $b_{i+1} = b_i + \varepsilon = \hat{b}_{i+1} + 2\varepsilon$): this way, he would obtain the higher slot, and hence higher CTR, with almost no change in the price he pays. But then bidder i, who had originally lowered his bid in order to lower his payment for the i-th slot, is now out-bid by $i+1$, and drops one position down. At this point, bidder i has an incentive to increase his bid again so as to re-gain his original position. Thus, the initial phase in which bidders start lowering their bids so as to lower their payments, given the original allocation, is followed by a phase in which bids are subject to an upward pressure, in an attempt to maintain the original position.

But since the higher valuation bidders have a higher willingness to pay for any given slot, this race to the top eventually re-establishes the original ranking, and hence it leads back to the efficient allocation: a low valuation bidder would stop competing for any given slot earlier than a high valuation bidder would, and different bidders would drop out of the race in increasing order of their valuation. But once the race-to-the-top is over, and the efficient ranking of bidders is re-established, then we are back to the original situation: holding positions constant, each bidder who obtains a slot has an incentive to decrease his own bid. And so it happens, until bids are so low that the race-to-the-top begins once again, and so on. Thus, because of the property of the GFP auction that bidders pay their own bid, no deterministic profile of bids $\hat{b} = (\hat{b}_1, \hat{b}_2, ..., \hat{b}_n)$ can form an 'equilibrium' of this auction, when bidders' valuations are stable.

In summary, when there is uncertainty over bidders' valuations, then the GFP auction admits an equilibrium which induces efficient allocations, just as in the baseline first-price auction with a single good. This is because the uncertainty over others' valuations translates into uncertainty over their bids, which in turn prevents bidders from lowering their bids without risking their slot. However, when there is no uncertainty over valuations, then the GFP has no 'pure strategy' equilibrium: the only equilibria must involve some randomization (if there is no uncertainty in valuations, then such randomization must be directly in the bids placed by the advertisers).[12]

The ultimate reason why the GFP ended up inducing bidding cycles was therefore that, for many keywords-auctions, the set of bidders and their valuations didn't present sufficient uncertainty to prevent the advertisers from engaging in the mechanism described above. The incentives to lower their bids were too strong, which in turn triggered the following reaction of aggressive bidding, and hence the cycle.

[12] See, for instance, Edelman and Schwarz (2007), which first provided this explanation for the shortcomings of the GFP auction.

But once bidders have gone over a few of these cycles, then they also understand that there isn't much of a point in triggering the race-to-the-top. It soon becomes clear that any such price war is doomed to be won by the higher valuation advertisers, and hence re-establish the original allocation, just with higher prices for everyone. Hence, after a few of such bidding cycles, advertisers realize that raising each others' bids in order to alter the final allocation is a desperate attempt. They would thus stop doing that, and accept instead the same allocation at the low bidding profile. This way, the bidding cycles generated by the lack of pure equilibria in the GFP auction favored an indirect form of collusion among the advertisers, which in turn eroded the revenues generated by the GFP auction.

2. The Google Revolution and the GSP-Auction

The phenomenon of bidding cycles observed in the GFP auction, which can be explained by its lack of pure equilibria, has been taken to be the main responsible for the ultimate abandonment of the GSP format, and for the creation of a new auction format, which would soon dominate this market: the Generalized Second Price (GSP) auction.

In February 2002, Google introduced the GSP auction as part of its *AdWords Select* bidding platform. Key to Google's success was the ability to incorporate advertisement in the clean layout of its pages, without diluting the informative content for the consumers. In the seminal paper which marked the birth of Google, its founders Sergey Brin and Larry Page complain that earlier advertising-funded search engines were "inherently biased towards the advertisers and away from the needs of consumers" (Brin and Page, 1998), which they deemed a major pitfall. The concern for building and maintaining a long-lasting consumer base is a central concern in Google's history, and can be explained in terms of the discussion from Section II.2.1. on how quality of the webpage can increase the value for the advertisers, and hence the profitability of Google's search pages.

But as we will discuss shortly, the strategic structure of the GSP auction and the simplicity of its rules turned out to be fundamental to ensure stable bidding behavior, and hence a solid revenue base, which boosted Google's business in an unprecedented way: on August 19th, 2004, Google went public with a valuation of $27 billion. In 2011, the company registered $37.9 billion in global revenues, of which $36.5 billion (96%) were attributed to advertising.[13] Google is now worth close to $300 billion.

[13] Source: Google Inc., from Blake, Nosko and Tadelis (2015).

Google's success turned the GSP into the mechanism of choice of all other major search engines, including earlier incumbent Yahoo!, its subsequent partner Microsoft-Bing, and Taobao in China. The GSP's supremacy among online ad auctions went essentially undisputed, until recently, when another major player in the industry attempted an alternative route, which we will discuss in Section III.3.

2.1. Economic Analysis of the GSP Auction

To understand the reasons of the GSP's success, it is useful to recall its rules. We begin with a presentation of the GSP that ignores the so called "quality scores" which essentially represent a re-weighting of bids by how strong of a match the advertiser is for the given search query. While being an integral part of the innovations introduced by Google's sponsored search auctions, quality scores are not intrinsically part of the GSP and, indeed, Taobao does not use quality scores in its GSP.

> In the *Generalized Second-Price (GSP) Auction,* bidders submit bids simultaneously. Bids are ranked from the highest to the lowest, the j-th highest bidder gets the j-th slot, and pays a price-per-click equal to the next (the $j+1$-th) highest bid.

Note that, given these rules, the GSP auction shares a very important property of the baseline (single-good) second-price auction. Namely, a bidder's bid determines which slot he gets, if any, but not the price-per-click he pays for that slot.

This means that, given a particular profile of bids $\hat{b} = (\hat{b}_1, ..., \hat{b}_n)$, we may have two cases: either (i) there is no bidder who, taking as given the others' bids, has an incentive to change his own (what economists call an *equilibrium*); or (ii) there is some bidder who would rather occupy a different slot (either a lower one –by lowering his bid below some of the lower ones– or a higher slot –by increasing his bid above some of the higher ones). The difference with respect to the GFP auction is that it would never be the case that a bidder would want to change his own bid but not his own position. This property is important because, if there is a profile of bids $\hat{b}$ such that every bidder prefers exactly the position which he obtains, given the resulting prices, then he would have no incentive to change his own bid. Hence, this basic property of the GSP auction overcomes the very basic problem underlying the bidding cycles in the GFP auction. Namely, the incentives bidders had in the GFP auction to lower their bids, holding the allocation constant.

The discussion thus far has focused on the similarities between the GSP and the baseline second-price auction for a single-object (we recall that the two are equivalent when there is a single slot on sale, $S = 1$). But when there is more than one object on sale, $S > 1$, there are also important differences between the two. In particular, in the GSP auction, everybody bidding truthfully (that is, setting $b_i = v_i$ for each i) is not an equilibrium anymore. Hence, in the GSP auction, bidding truthfully is not a 'dominant strategy' the way it was in the baseline second-price auction.

To see this, suppose that there are four bidders, with valuations $v_1 = 3$, $v_2 = 2.9$, $v_3 = 2.8$, and $v_4 = 0.1$, and suppose that there are three slots on sale, with CTRs $x_1 = 10$, $x_2 = 9$ and $x_3 = 8$, and that everyone is bidding truthfully. Then, bidder 1 obtains the highest slot at a price equal to v_2, and obtains a payoff equal to $x_1 \cdot (v_1 - v_2) = 1$. On the other hand, given the current bids, the price-per-click paid for the third slot is very low: it is equal to $v_4 = 0.1$. But, if rather than bidding truthfully and obtaining the best slot at a price very close to his own valuation, bidder i placed a low bid (say 0.5, or any other bid between v_4 and v_3), he would obtain the third slot at the very low price of v_4. This would result in a payoff of $x_3 \cdot (v_1 - v_4) = 23.2$. Hence, in this case, the highest valuation bidder would find it much more convenient to obtain the worst slot at a very low price, rather than bidding truthfully and obtaining the best slot at a very high price. But this shows that now bidding truthfully is not a dominant strategy anymore, and hence despite the similarities between the two auctions, the strategic behavior in the GSP auction is much more complex than in the baseline second-price auction.

Economic analysis shows that the GSP auction can have many equilibria, but one particular equilibrium has received a particular attention in the theoretical economics literature, and has become the benchmark to study the competitive equilibrium in the GSP auction. Besides its many theoretical advantages, one important reason why this particular equilibrium is especially interesting is that it conforms with the instructions provided by Google's *AdWord* tutorial on how to bid in the auctions.[14] In this equilibrium, (i) bids are ranked according to bidders' valuations (that is, $b_1 > b_2 > ... > b_n$, so that the resulting allocation is efficient); (ii) the lowest-valuation bidders who do not obtain a slot bid truthfully (that is, $b_i = v_i$ for all $i > S$); and (iii) all other bidders $i = 2, ..., n$ place a bid b_i which makes them indifferent between obtaining the i-th position at the current price (which is equal to the next highest bid, b_{i+1}) and climbing up one position (to CTRs x_{i-1}) paying a price-per-click equal to their own bid b_i (in math,

[14] The theoretical properties of this equilibrium were first studied by Varian (2007) and Edelmann, Ostrovsky and Schwarz (2007). For the Google AdWord tutorial in which Hal Varian teaches how to maximize profits by following this bidding strategy, see: http://www.youtube.com/watch?v=jRx7AMb6rZ0

this means that $b_i = v_i - \frac{x^i}{x^{i-1}}(v_i - b_{i+1})$ for all $i = 2, \ldots, S$); (iv) the bid of the top bidder is not uniquely pinned down, the only restriction being that its value exceeds that of the next bid.

To illustrate this competitive equilibrium, as well as other points in the subsequent discussion, we will repeatedly refer to the following example (the example is taken from Decarolis, Goldmanis and Penta, 2017):

Example 1. Consider an auction with four slots and five bidders, with the following valuations: $v_1 = 5$, $v_2 = 4$, $v_3 = 3$, $v_4 = 2$ and $v_5 = 1$. The CTRs for the four positions are the following: $x_1 = 20$, $x_2 = 10$, $x_3 = 5$, $x_4 = 2$. In this case, the *competitive equilibrium benchmark in the GSP auction* is as follows: $b_5 = 1$, $b_4 = 1.6$, $b_3 = 2.3$ and $b_2 = 3.15$. The highest bid $b_1 > b_2$ is not uniquely determined, but it does not affect the revenues because it doesn't affect the payment of the highest bidder (it only determines the fact that he gets the highest slot). In this example, the total revenues are 96, and the resulting allocation is clearly efficient.

In the discussion above we intentionally disregarded a feature that was prominently pushed through by Google when it launched its GSP model: quality scores. The main insight is that some advertisers might value appearing on keywords that are a poor match for their products with the logic of creating a potential "lead" (*i.e.,* building a name recognition that might generate future sales) at a very low price (a click on their link will be unlikely). This, however, would hurt the search engine both in the short run, through the low click-through-rate, and in the long run, as consumers using the search engine might find particularly annoying to be exposed to advertisements unrelated to their queries. To solve these problems, Google's version of the GSP ranks advertisers not only by their bids but by the product of their bids and a quality score. The latter is a function of past click behavior and, like the algorithm for Google's organic search results, assigns more weight to advertisers with a greater likelihood of being clicked. The mechanics of the auction with quality scores is nearly identical to what we illustrated above, but with a more involved notation. For that extension we therefore defer to our more technical study, Decarolis, Goldmanis and Penta (2017).

3. Facebook and the VCG Auction

Around 2007, Facebook began experimenting with the VCG for its own display ad auctions and, by 2015, its transition to this format for all its ad

auctions was complete. These display ad auctions are different from those of the search engines we discussed so far. That is because these auctions are not generated by keywords and because they raise specific challenges to integrate ads within Facebook's organic content. But these technicalities aside, they boil down to the same kind of economic problem we have been discussing all along: a multi-unit auction problem.

Before John Hegeman, an economics MA graduate from Stanford, took the role of Facebook's chief economist, the (multi-unit) VCG had had a limited impact outside of academia. Perhaps for this reason, or for the somewhat byzantine VCG payment rule, the industry's initial reaction to Facebook's innovation was one of surprise (*cf.* Wired, 2015). But Facebook and its VCG auction are now essential parts of this industry: in the second quarter of 2015, Facebook pulled in $4.04 billion and, together with Twitter, it has become one of the largest players in display ad auctions. According to Varian and Harris (2014), around 2012 also Google considered a transition to the VCG auction for its search auctions, but ultimately decided to switch to VCG exclusively for its contextual ads sales, because of the perceived risks associated with communicating to bidders the complex VCG payment rule.

The VCG is a classic and widely studied auction in the academic literature that involves a fairly complex payment scheme. As we will explain in Section III.3.1., it is designed to price the externalities that each bidder forces on others in the efficient allocation. On the other hand, as we will also discuss in Section III.3.1., the VCG has the advantage that bidding truthfully is a dominant strategy, just as in the baseline second-price auction. The resulting allocation therefore is efficient. The GSP auction in contrast has very simple rules (the k-highest bidder obtains the k-highest slot at a price-per-click equal to the $(k+1)$-highest bid), but it gives rise to more complex strategic interactions. The relative merits of the two auctions therefore appear unclear, at least at first glance.

However, consider once more our earlier auction problem from Example 1:

Example 2. There are four slots and five bidders, with the following valuations: $v_1 = 5$, $v_2 = 4$, $v_3 = 3$, $v_4 = 2$ and $v_5 = 1$. The CTRs for the four positions are the following: $x_1 = 20$, $x_2 = 10$, $x_3 = 5$, $x_4 = 2$. But this time suppose that the seller uses a VCG auction, rather than the GSP. As we will discuss shortly, bidding truthful is a dominant strategy in the VCG. In this equilibrium, everybody bids $b_i = v_i$, and hence the resulting allocation is the same as the GSP auction. Moreover, applying the formula for the VCG payments, it is easy to check that the total revenues are exactly the same which would be obtained in the benchmark competitive equilibrium of GSP auction: 96.

Hence, based on this example, it seems that the GSP auction is both simpler and ensures the same revenues and allocation as the VCG: while the increased complexity of the VCG ensures that bidding truthfully is a dominant strategy, it does not seem to yield higher revenues in this setting, nor a better allocation.

Economic theorists have shown that this outcome-equivalence result between the VCG auction and the benchmark competitive equilibrium of the GSP auction holds in general (See Edelmann, Ostrovsky and Schwarz, 2007). Combined with the simplicity of the GSP rules, this result has provided a rationale for the GSP's striking success and, until recently, its almost universal diffusion.

The next subsection provides a more in-depth look at the VCG auction, and its relation with the GSP and the baseline second-price auction. An attempt to explaining why the VCG might be actually preferable to the GSP is provided in Section IV, in which we discuss further recent trends in the market, which operate along with the changes in the auction formats and affected their performance.

3.1. Economic Analysis of the VCG Auction

We begin by explaining why bidding truthfully is a dominant strategy in the VCG auction. To this end, we recall the rules of this auction:

- In the *(VCG) Auction,* the k-th highest bidder gets the k-th slot, and pays a price equal to a weighted sum of all lower bids, where weight of the l-th highest bid (for $l > k$) is equal to $(x_{l-1} - x_l)$, where we set $x_k = 0$ for all $k > S$.

First note that, similar to both the GSP and the baseline (single-unit) second price auction, each bidder's own bid does not affect directly the price he pays for the slot he obtains (besides determining which slot he gets). If i places the k-highest bid, he obtains the k-th slot, and pays a price which only depends on the lower bids (each weighted by the term $(x_{l-1} - x_l)$ for all $l > k$). It is thus clear that, unlike the GFP auction, bidders in the VCG wouldn't have a strict incentive to lower their bids, holding the allocation constant. In fact, when there is a single-object on sale ($S = 1$), then the VCG coincides with the baseline second-price auction, just like the GSP does.

To see that it would never be optimal to bid more than one's own valuation, note that (similar to the baseline second-price auction), bidding $b_i > v_i$ would

only affect the outcome in the event that some of the other bids were above v_i. But, in that case, the gain due to the increased CTR would be more than offset by the higher price: suppose that, by bidding truthfully, agent i obtained position k, whereas by bidding $b_i > v_i$ he climbed up one position, to slot $k-1$. Then, this means that there exists exactly one opponent, say j, whose bid b_j is such that $v_i < b_j < b_i$. Now, bidder i's increase in utility due to climbing one position up from k to $k-1$ is equal to $(x_{k-1} - x_k) \cdot v_i$. But the increase in price is equal to $(x_{k-1} - x_k) \cdot b_j$, since now bidder j has fallen below bidder i, increasing his payment. But note that, by assumption, $b_j > v_i$ in this case, and hence the increase in payment is larger than the increase in payoff due to the higher slot.

Increasing one's bid above one's own valuation in order to climb one position up therefore would never be optimal. A similar argument applies to the case in which bidding $b_i > v_i$ allows bidder i to climb more than one position up. In all these cases, increasing one's bid above one's own valuation either has no effect on the ultimate allocation, or it lowers the overall payoff, since it induces an increase in payment higher than the increase in utility due to obtaining a better slot. A symmetric argument also shows that lowering one's bid below one's own valuation never increases the payoff: it either has no effect on the resulting allocation and payoffs, or it induces a lower slot in a suboptimal way, in that climbing up to the original slot would induce an increase in utility which is larger than the increase in payment it is associated with.

In conclusion, exactly like in the baseline second-price auction, bidding truthfully is an optimal strategy in the VCG regardless of what others do. Recall that this was not the case in the GSP auction, in which in fact bidding truthfully was not an equilibrium (see Example 1). In this sense, the VCG truly is the correct way of generalizing the properties of the baseline second-price auction to the case in which multiple objects are on sale. Despite the seemingly closer connection between the GSP and the baseline second-price auction, the GSP has very different properties from it. Those properties are instead inherited by the more complicated VCG auction: bidding truthfully is dominant, and it induces an efficient allocation.

This is not by chance. In fact, academic economists designed the VCG auction and its generalizations precisely to achieve these goals. These ideas have been applied for instance to ensure socially efficient outcomes not only in auctions, but also in environmental economics, or for solving the problem of optimal provision of public goods. The key idea behind the VCG payments, and the reason why they induce efficient allocations, is that they provide a sophisticated way of pricing the externalities which may otherwise induce

inefficiencies, very much like the *Pigouvian* taxes used to reduce firms' polluting emissions.

To see this, note that if everybody bids truthfully in the VCG auction, then each bidder i obtains the i-th slot, and pays a price equal to a weighted sum of the valuations of all agents $j > i$, where each v_j is weighted by the term $(x_{j-1} - x_j)$. Formally, the payment for the i-th position is equal to $\sum_{j=i+1}^{n} (x_{j-i} - x_j) \cdot v_j$. In other words, bidder i pays for the i-th slot the total value of the *externality* that he imposes on others. To see that this is actually the case, it is useful to pause for a moment and consider what is i's externality on others: if bidder i and his bid were removed from the system, then the bidders with valuation higher than i (that is, those indexed with $j < i$) would still obtain the same slots. However, if i and his bid were removed from the auction, then all bidders below him (the j's such that $j > i$) would each climb up one position. Hence, each j would move from CTR x_j to CTR x_{j-1}. The expected gain in utility for such j is thus $(x_{j-1} - x_j) \cdot v_j$. Hence, the total externality that i's presence forces on others is that it prevents all bidders with lower valuation to each climb up on position in the ranking of slots, which displaces a utility of $(x_{j-1} - x_j) \cdot v_j$ for each $j > i$. The total externality of agent i in slot i therefore is precisely $\sum_{j=i+1}^{n} (x_{j-i} - x_j) \cdot v_j$, which is the VCG payment for the i-th slot if everybody bids truthfully.

IV. RECENT DEVELOPMENTS: NEW PLAYERS AND AGENCY BIDDING

Alongside the evolution of auction platforms, this market has witnessed profound changes on the advertisers' side as well. In the early days of online ad auctions, advertisers bid through their own individual accounts. Moreover, these accounts were often managed separately across different bidding platforms. But already back in 2011, a large share of advertisers in the US delegated their bidding activities to specialized digital marketing agencies (DMAs): A survey by the Association of National Advertisers of 74 large U.S. advertisers indicates that about 77% of the respondents in 2011 fully outsourced their search engine marketing activities (and 16% partially outsource them) to specialized agencies.[15] Analogously, a different survey of 325 mid-size advertisers by Econsultancy reveals that the fraction of companies not performing their paid-search marketing in house increased from 53% to 62% between 2010 and 2011.[16] Moreover, many of these DMAs belong to a handful of networks (seven in the U.S.) that conduct all bidding activities through centralized agency trading desks (ATDs). As a result, with increasing frequency, the same entity (be

[15] Source: ANA (2011).

[16] Source: Econsultancy (2011).

TABLE 1

CPC IS THE AVERAGE COST-PER-CLICK IN $US. VOLUME IS THE NUMBER OF MONTHLY SEARCHES, IN THOUSANDS. POSITION REFERS TO RANK AMONG PAID SEARCH LINKS ON GOOGLE'S RESULTS PAGE FOR THE RELEVANT KEYWORD

Keyword	*CPC*	*Volume*	*Position*	
			Habitat	Salv. Army
Habitat for humanity donations pick up	4.01	40	1	4
Charities to donate furniture	1.08	20	3	9
Donate online charity	0.93	20	11	10
Website for charity donations	0.90	19	11	6
Salvation army disaster relief fund	0.03	20	2	1
Giving to charities	0.05	30	8	5

Source: 2016 US Google sponsored search data from SEMrush, in Decarolis, Goldmanis and Penta (2017).

it DMA or ATD) bids in the same auction on behalf of different advertisers, a phenomenon we label as "common agency."[17]

As we will argue, this recent market trend is bound to alter the very functioning of the main auction formats, and has thus the potential to shake up the entire industry. It also creates new opportunities for marketing agencies to create surplus for their clients. The reason is that this issue of common agency clearly changes the strategic interaction, as these agencies now have the opportunity to lower their payments by coordinating the bids of their clients.

1. The Phenomenon of Common Agency

The case of Merkle, one of the major agencies in the U.S., provides a clear example of the common agency phenomenon we introduced above. A quick visit to Merkle's website immediately reveals that many of Merkle's clients operate in the same industries, and are therefore likely to bid on the same keywords.[18] For instance, data from Redbook (the leading public database to link advertisers to their agencies) confirm that Merkle managed the campaigns

[17] Another form of common agency also common in the retailing sector involves the fact that, since both brands and retailers can advertise on the same keywords, it is common for manufacturers to coordinate with their retailers on search ad spending. See Cao and Ke (2017) for an analysis of this form of cooperative advertising.

[18] See: https://www.merkleinc.com/who-we-are-performance-marketing-agency/our-clients

of many competing advertisers. This is, for instance, the case of two leading charities, *Habitat for Humanity* and *Salvation Army,* both of which in 2016 were bidding through Merkle in the same auctions for hundreds of keywords. Out of all these keywords, Table 1 reports the six with the highest search volume specifying for each of them the average cost-perc-click and positions of both *Habitat for Humanity* and *Salvation Army.* Similar examples can be identified for nearly every industry: for clothing, *Urban Outfitters* and *Eddie Bauer* use Rimm-Kaufman; for pharmaceuticals, *Pfizer* and *Sanofi* use Digitas; etc. Table 1 reports the top six of these keywords, in terms of their average cost-per-click (CPC).

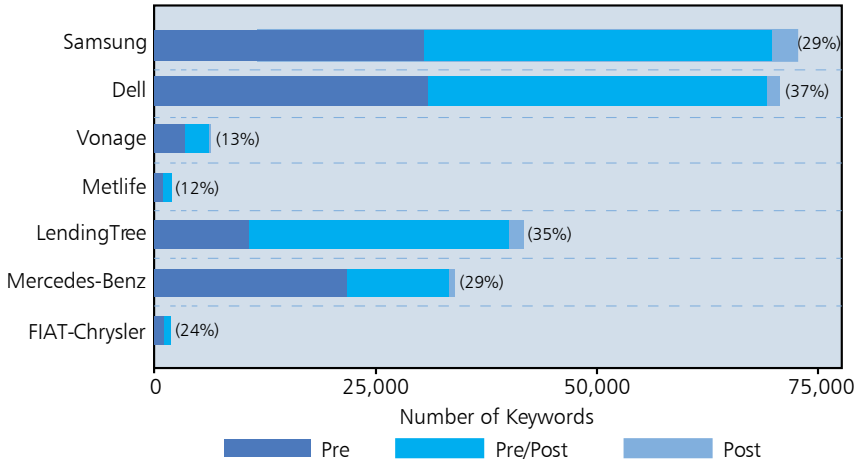
The common agency problem is made even more relevant by yet another recent phenomenon, the formation of 'agency trading desks' (ATDs). While several hundred DMAs are active in the US, most of them belong to one of the seven main agency networks (Aegis-Dentsu, Publicis Groupe, IPG, Omnicom Group, WPP/Group M, Havas, MDC), which operate through their corresponding ATDs (respectively: Amnet, Vivaki, Cadreon, Accuen, Xaxis, Affiperf and Varick Media). ATDs' importance is growing alongside another trend in this industry, in which DMAs also play a central role. That is, the ongoing shift towards the so called 'programmatic' or 'algorithmic' real time bidding: the algorithmic purchase of ad space in real time over all biddable platforms through specialized software. ATDs are the units that centralize all bidding activities within a network for 'biddable' media like Google, Bing, Twitter, iAd, and Facebook. Hence, while DMAs were originally not much more sophisticated than individual advertisers, over time they evolved into more and more sophisticated players, and their diffusion and integration through ATDs has made the issue of common agency increasingly frequent.

Below we will discuss the implications that common agency may have in terms of inducing collusive bidding strategies in the various auction formats. But this need not be the only way in which agencies implement coordinated strategies. One alternative could be to split the keywords among an agency's clients, so that they do not compete in the same auctions. This 'bid retention' strategy is obviously advantageous in single-unit auctions, but in principle it might be used in multi-unit auctions too. A recent episode, also part of the trend towards concentrated bidding outlined above, may help us illustrate the significance of the potential for bid coordination which we hinted at above.

In July 2016, Aegis-Dentsu acquired Merkle, which was not previously affiliated to any network. At that time, many of Merkle's clients were bidding on the same keywords as some of Aegis-Dentsu's advertisers. For instance, in the electronics sector, *Dell* and *Samsung* were in Merkle's portfolio, placing bids on keywords also targeted by Aegis-Dentsu's clients *Apple, HP, IBM/Lenovo* and *Intel*. Other examples include: in the financial sector, Merkle's *Lending Tree*

FIGURE 1

NUMBER OF KEYWORDS ON WHICH EACH OF MERKLE'S ADVERTISERS BIDS ALONGSIDE AT LEAST ONE MEMBER OF THE AEGIS-DENTSU NETWORK
As a share of the total number of keywords on which it bid, in parenthesis, between June 2015 and January 2017



Notes: Merkle's acquisition by Aegis-Dentsu was in July 2016. The graph shows whether bids on these 'shared' keywords occurred only pre-acquisition (dark blue: all keywords appearing only before July 2016), only post-acquisition (turquoise: all keywords appearing only after July 2016), or both pre- and post-acquisition (blue: all keywords appearing both before and after July 2016).

Source: Decarolis, Goldmanis and Penta (2017) using keyword-level data provided by SEMrush.

and *Metlife* were bidding in auctions alongside Aegis-Dentsu's *Capitalone, Discover, Fidelity, Equifax, JP Morgan-Chase;* for car manufacturers, Merkle's *FIAT-Chrysler* and *Mercedes-Benz USA* bid alongside Aegis-Dentsu's *Toyota, Volkswagen, Subaru;* in phone services, Merkle's *Vonage* bid alongside Aegis-Dentsu's *T-Mobile*.[19]

This acquisition therefore further increased the potential for coordinated bidding. Figure 1 reports, for each of Merkle's advertisers listed above, the fraction of the total keywords on which they were bidding at the same time as some of Aegis-Dentsu's clients, and whether joint targeting of such keywords happened only pre-acquisition, only post-acquisition, or both pre- and post-acquisition. Although there is some variation among these advertisers, we clearly see that shared keywords are a quantitatively large phenomenon also post-acquisition (interestingly, a small fraction of keywords are shared *only* post-acquisition). Hence, this case suggests that coordinated bidding through a

[19] Source: Redbook.

common agency in the same auction is a relevant phenomenon. Clearly, the figure also suggests that keyword split among the advertisers can be important. Depending on the relevance of the keywords for the different advertisers and on the easiness of splitting markets, we can expect both phenomena to characterize agency bidding strategies.

2. Agencies' Opportunities and their Potential Consequences

To understand the potential impact that agencies may have on online auctions, consider the VCG auction first. As discussed above, when advertisers compete with each other, bidding truthfully is a dominant strategy. Moreover, if everybody follows this strategy, slots are allocated efficiently to bidders and each bidder pays a price equal to the externality he forces on others. But now suppose that a single agency controls the bids of two advertisers, say the ones with the highest and third-highest valuation, while all other advertisers are still bidding independently. Then, it is still the case that bidding truthfully is a dominant strategy for the independent bidders. However, the agency now has an incentive to lower the bid of her lower member (the third): that is because, given the rule which determines the VCG payments, an advertiser's bid directly affects the payment of all advertisers placing bids above him. So, by lowering the bid of her lower member(s), the agency would lower the payments of her higher member, with no need to alter the resulting allocation at all. Note that this opportunity arises solely because the agency controls the bids of different advertisers in the same auction: if an agency's clients competed in different auctions, then there would be no opportunity to lower the payments of its clients through manipulation of their bids.

At a minimum, this observation points at a new opportunity that agencies have to generate surplus for their clients –besides other activities aimed at improving their advertisement strategies, overall appeal of the product, and so on. That is, DMAs have now the opportunity to generate surplus by manipulating the bids of their clients in order to reduce their payments in the VCG auction, and hence ultimately their cost for online advertisement. This has important consequences for both the agencies and the auction platforms.

From the agencies' viewpoint, this raises questions on (i) what is the optimal strategy to generate surplus through coordinated bidding in a given auction, for a given portfolio of clients; and (ii) what is the optimal composition of the portfolio of clients in order to increase the agency's ability to generate surplus through coordinated bidding. The analysis of these points is rather complicated, but we can discuss here the main trade-offs.

For what concerns point (i) –the optimal coordinated bidding strategy in a given auction– note that the argument above suggests that the high valuation clients of a DMA gain more as the bids of the low-valuation clients are kept as low as possible. Of course, however, agencies cannot just lower the bids of their lower clients as much as possible, since they still need to ensure that these clients are sufficiently satisfied with the allocation they get and price they pay that they would not decide to abandon the agency.

Hence, this situation of common agency in an auction requires solving a conflict of interest between the agency's clients with higher valuations and those with lower valuations. The solution of the optimal trade-off between these opposing interests is fairly easy to solve for the case of the VCG auction, but it is much more complicated for the GSP auction. It is clear, however, that given the complex formula to determine the VCG payments, agencies' margin to generate extra surplus by coordinating the bids of different advertisers are very large. This is the case even when the agency controls a very small number of bidders. The next example illustrates the point numerically in our running example:

Example 3. Consider the environment in Example 1, and suppose that the agency controls both the first and the third bidder, 1 and 3. Now, suppose that the agency lowers b_3 to the minimum level which still esures he maintains the third position. Then, this has no effect on the slot and payment of the third bidder, but it decreases the payment of both the second bidder (who does not belong to the agency), and of the first. Their payments decrease by 5 each. Hence, by manipulating bids in this way, the agency is able to generate an extra surplus of 5 for her highest member, at absolutely no cost for her lower member. Note that, given this particular numerical example, an increase of payoff of 5 entails a non-trivial percentage of the overall payoff. So this alone suggests that the effects may be very sizeable, even if the agency only controls two bidders, and without necessarily harming any of her clients.

For what concerns point (ii) –which composition of the portfolio of clients maximizes an agency's ability to generate surplus in this way– it seems obvious that the more the agency bidders in the same auction, the greater the agency's ability to generate surplus through coordinated bidding. But apart from these obvious considerations, the general answer is more complex. For instance, holding constant the total number of clients that an agency controls in the same auction, is it better for her to have clients with high or with low valuations? The surprising, general answer in this case is neither: what really matters to boost an agency's ability to extra surplus through coordinated bidding is not so much the level of her clients' valuations, but the specific position they occupy relative to the independents, and the exact values of the CTRs. In the example above,

for instance, if the agency controls the highest overall bidder, then the agency's ability to generate surplus is maximized by controlling the second bidder. But in an auction with different CTRs it may be the third, or the fourth.

It is clear, however, that holding everything else constant, the agency's impact on the overall revenues of the auction are higher as her clients occupy *lower* positions in the ranking of valuations. That is because if the agency lowers the bid of a low bidder in the ranking of valuations, she is going to lower the payments of *all* bidders above her clients, whether or not they are the agency's clients or just independent bidders. Hence, the agency's ability to maximize the surplus she may generate for her clients in general does not coincide with her potential to harm the revenues of the auction platform.

Example 4. Note that, in the example above, while the agency can lower her clients' total payments by 5, the total revenue loss for the auction platform is 10: the total revenues with coordinated bidding is 86, as opposed to 96 of the competitive benchmark.

These observations suffice to cast serious doubts on the revenue properties of the VCG auction in the presence of coordinated bidding: since the bid of any bidder affects the payments of all bidders placing bid above his, even small bid manipulations may have strong effects on revenues. In this sense, the VCG auction seems very vulnerable to the agencies' potential for coordinating the bids of their clients.

In contrast, the GSP payments are such that a bidder's bid only affects the payment of the advertiser who places a bid immediately above his. Based on this observation, one is tempted to conclude that the GSP auction would be more resilient to the presence of agency bidding. This conclusion, however, overshadows the complexity of the strategic interaction generated by the GSP auction. In particular, since in this auction the independent bidders do not have a dominant strategy, it may be that the agency's manipulation of clients' bids may indirectly affect also the bids of the independents. If this is the case, then the resulting effects on the agency's payments and on the platform's revenues are unclear.

In fact, as we will explain shortly, we have reasons to believe that the GSP may potentially be even more fragile than the VCG auction. The source of the GSP's fragility, and the complexity of agency bidding in this context, can be understood thinking about an agency that controls the first, second, and fourth highest bidders in an auction. The agency in this case can lower the highest bidder's payment by lowering the bid of the second, without necessarily

affecting either his position or his payment.[20] Given the rules of the GSP auction, the agency can benefit from this simple strategy only if two of her members occupy adjacent positions. But due to the GSP's complex equilibrium effects, the agency can do more than that. For instance, suppose that this agency shades the bid of her lowest member, with no direct impact on her other clients' payments. Intuitively, if this bid is kept persistently lower, then the logic of independent bidders' behavior in the competitive equilibrium benchmark suggests that the third highest bidder, who is an independent, would eventually lower his bid. But not only would this lower the second bidder's payment, it would also give the agency extra leeway to lower the second-highest bid, to the greater benefit of the highest bidder. Revenues in this case diminish for both the *direct effect* (lowering the 2-nd highest bid lowers the highest bidder's payment) and for the *indirect effect* (lowering the 4-th highest bid induces a lower bid for the independent, which in turn lowers the second bidder's payment). Hence, even an agency controlling a small group of advertisers may have a large impact on total revenues. The next example illustrates how this mechanism works in the context of our running example:

Example 5. Consider again the environment of Example 4, in which the agency controls the first and third bidder, but now suppose that the platform adopts the GSP auction format. Now, suppose that the agency lowers the bid of the third highest bidder almost all the way down to 1.6, the competitive equilibrium bid of the fourth bidder, who is not controlled by the agency. Then, both the position and the payments of the third bidder are not affected. Yet, applying the logic of competitive bidding to the second bidder (who is not controlled by the agency), he would lower his bid from 3.15 (see Example 1) to 2.8. This in turn lowers, indirectly, the payment of the highest bidder, who is an agency client. Overall, the total revenues in this configuration are 82, which are lower than in the VCG auction with the same agency structure (86), and of course lower than the competitive benchmark, which generated revenues of 96 in both auctions.

The basic insight that the GSP is more vulnerable to coordinated bidding from an agency has more general validity. The example also suggests that the problem of identifying the optimal bidding strategy for the agency, as well as the optimal composition of the portfolio of its client, is much more complicated in the GSP than in the VCG auction. For instance, in the example above, one may wonder if the agency could push the bid of her lower member (bidder 3) further down. The problem there is that then the next independent bidder might have

[20] Clearly, we are implicitly assuming that an agency has an incentive to lower its clients payments, for a given amount of clicks. This is indeed the case since the typical arrangement in the agency-advertiser relationship entails the agency receiving a flat fee per ad campaign, so that an agency's probability of future contracts derives from its ability to generate value for the advertiser, for instance by achieving cost-per-click savings.

TABLE 2

SUMMARY OF RESULTS IN EXAMPLES

Valuations	*Competitive VCG*	*Competitive GSP*	*VCG with agency*	*GSP with agency*
5	5	b_1	b_1	b_1
4	4	3.15	4	2.8
3	3	2.3	2^+	1.6^+
2	2	1.6	2	1.6
1	1	1	1	1
Revenues	96	96	86	82

Sources: Summary of results in Examples 1-5. Agency clients' bids and valuations are in bold. Numerical examples taken form Decarolis, Goldmanis and Penta (2017).

an incentive to raise his bid and climb up one position, hurting bidder 3 who may at that point decide to abandon the agency. In some auctions (that is, depending on the CTRs and on the valuations of the bidders), it may not be sustainable for the agency to induce inefficient allocations, but in other auctions it can be. The exact optimality therefore requires a correct understanding of the strategic reaction of the independent bidders and of the payoff implications.

In the opposite direction, one may worry that this kind of behavior from the agency might be detected as collusive, and possibly be punished by an external observer (for instance, a public authority or by the auction platform itself). If one wanted to address these concerns, then the optimal strategy of the agency would be less aggressive in lowering the bidder 3's bid in the previous example.[21] Hence, while the agency has ample margins to generate surplus for their clients through coordinated bidding, the optimal agency bidding strategy in the GSP requires a sophisticated analysis of the strategic interactions it generates.

On the other hand, it seems clear that the GSP may be more vulnerable to agencies' exploitation of these opportunities, than the VCG auction might be. This is a strong statement because the VCG auction is well-known to be highly susceptible to collusion, but it is especially noteworthy if one considers the sheer size of transactions currently occurring under the GSP. It also suggests a rationale for why Facebook's recent adoption of the VCG mechanism was so successful, despite the early surprise it provoked, and for why the last few years have recorded a steady decline in ad prices.[22] Google, for instance, reports

[21] We address these complex issues in Decarolis, Golmanis and Penta (2017).

[22] On the early surprise that Facebook adoption of the VCG auction generated, see Wired (2015).

passing from a positive growth rate in its average cost-per-click of about 4 percent per year in the four years before 2012, to a negative growth rate in each year since then, with an average yearly decline of 9 percent.[23]

The striking fragility of the widespread GSP auction we briefly discussed in this chapter suggests that further changes are likely to occur in this industry, raising important questions from different perspectives. These include (i) new opportunities for digital marketing agencies to generate surplus for their clients; (ii) novel issues for the existing auction platforms, and novel challenges to improve the design of the main auction formats; (iii) potential implications for antitrust authorities and for the consumers' welfare.

Since we already discussed the first two points, we conclude this section commenting on the latter. The optimal bidding strategies for the agencies we described above share important features with the behavior of collusive buying consortia, which have been sanctioned in the past by antitrust authorities.[24] One may thus be tempted to conclude that our similar behavior from the agencies' part might be sanctioned in a similar way. However, the specificities of the market suggest a more nuanced view of the harm to consumers. We return to this point in the concluding remarks in the next section.

V. A LOOK AHEAD AND CONCLUDING REMARKS

One interesting open question is whether the concerns discussed above may or may not be mitigated by competition between agencies. Although multiple agencies each with multiple bidders in the same auction seem rare at the moment (this is largely due to the agencies' specialization by industry), the question is nonetheless relevant because the phenomenon may become more common in the future. If an increase in agency competition restored the good properties of these auctions, then the diffusion of marketing agencies need not lead to major structural changes in this industry.

While only the evidence will tell, economic theory offers arguments to be skeptical of the healing potential of competition between agencies in this setting. As we formally show in Decarolis, Goldmanis and Penta (2017), for certain agency structures, agency competition mitigates the revenue losses in both the GSP and VCG auctions just as one would expect; but for other agency structures, agency competition has a particularly perverse impact on both

[23] Source: 10-k filings of Alphabet inc.

[24] See, for instance, the case of the tobacco manufacturers consortium buying in the tobacco leaves auctions, *United States v. American Tobacco Company, 221 U.S. 106 (1911).*

auction formats. That is because, from the viewpoint of an agency bidding for multiple clients, these auction mechanisms have a flavor of a first-price auction: even holding positions constant, the total price for an agency's client (except its lowest placed bidder) depends on the bids placed by the agency itself. With multiple agencies, this feature of agency bidding may lead to non-existence of pure equilibria, very much like the case of competitive (non-agency) bidding in the GFP auction. But as seen in the early days of this industry, when the GFP was adopted, lack of pure equilibria may generate bidding cycles which eventually lead to a different form of collusive outcomes and low revenues. As we discussed in Section III.1, these bidding cycles are one of the primary causes for the transition, in the early '00s, from the GFP to the GSP auction. Hence, not only does agency competition not solve the problems with these auctions, but it appears likely to exacerbate them, giving further reasons to expect fundamental changes in this industry.

As we pointed out earlier, the phenomenon of common agency opens new opportunities for digital marketing agencies to generate surplus for their clients, by both improving their bidding strategies in the existing online auction formats, and to structure the composition of their portfolio of clients in order to maximize their ability to manipulate the prices paid in these auctions. The optimal strategies are very complex to determine, especially for the GSP auction, as they require a careful understanding of the strategic interaction generated by these auction formats. It is clear however that the potential impact on agencies' profits and on auction platforms revenues are huge, and may have the potential to disrupt the current market arrangement and especially the prevailing auction formats.

As we also mentioned, it would be sensible for agencies to be cautious in manipulating the bids of their clients, as they may be concerned that their behavior may be detected as collusive, and possibly be punished by an external observer (for instance, a public authority or by the auction platform itself). In Decarolis, Goldmanis and Penta (2017), however, we show that even imposing an *undetectability constraint,* the optimal strategies for the agencies may significantly reduce their clients' payments, and hence extract suplus from the auction.[25]

All these issues are in fact potentially relevant from an antitrust perspective. In many ways, agency behavior in our model is analogous to that of buying consortia, which have been sanctioned in the past (*see United States v. American Tobacco Company, 221 U.S. 106 (1911)*). Nevertheless, the specificities of online ad market suggest a more nuanced view of the harm to the consumers. First,

[25] See Decarolis, Goldmanis and Penta (2017) for further details.

although our discussion focuses on agencies' role to coordinate their clients' bids, agencies in this market have other roles which are expected to improve the efficiency of the system (*e.g.,* in improving sellers' ability to reach new consumers, improving advertisers' campaigns, bringing new advertisers to the market, etc.) Second, it is likely that the degree of competition between different search engines is substantially less than that between most of advertisers. Since the lower auction prices due to agency bidding imply a reduction in the marginal cost advertisers pay to reach consumers, advertiser competition implies that some savings are passed on to consumers. Therefore, harm to consumers would result only if the agency engages in coordinating not only auction bids, but also the prices charged to consumers. Third, bid coordination can negatively affect the quality of the service received by consumers by further exacerbating the advantage of dominant search engines relative to fringe ones. In Europe, for instance, where 90% of the searches pass via Google, agencies might be rather careful not to harm Google given the risk of being excluded from its results page. Smaller search engines cannot exert such a threat because agencies are essential to attract new customers. The shift of revenues from small search engines to marketing agencies could thus deprive the former of the essential resources needed for technology investments. Thus, to the extent that competing search engines exert pressure for quality improvements, bid coordination poses a threat to consumer welfare. Quality of the links is indeed considered relevant for antitrust actions. For instance, in the Google case before the European antitrust authority, the Commission decided to fine Google 2.42 billion euro for abusing dominance as search engine by giving illegal advantage to own comparison shopping service, presenting links of inferior quality aimed at directing consumers to Google's own outlets.[26]

BIBLIOGRAPHY

Association of National Advertisers (ANA) (2011), *Trends in Digital Agency Compensation,* 4th Edition, ANA publishing.

Blake, T.; Nosko, C., and S. Tadelis (2015), "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment," *Econometrica,* 83(1): 155-174.

Cao X., and T. T. Ke (2016), "Cooperative Search Advertising," MIT Sloan School, *Working Paper,* 5176-16.

Decarolis, F.; Goldmanis, M., and A. Penta (2017), "Marketing Agencies and and Collusive Bidding in Online Ad Auctions, *NBER working paper,* 23962.

[26] See European Commission, Antitrust/Cartel Case no. 39740 Google Search (Shopping).

Econconsultancy (EC) (2011), "State of Search Marketing Report 2011," accessed on January 2012 at: http://econsultancy.com/us/reports/sempo-state-of-search

Edelman, B.; Ostrovsky, M., 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.

Edelman, B., and M. Ostrovsky (2007), "Strategic Bidder Behavior in Sponsored Search Auctions," *Journal of Decision Support Systems,* v.43(1): 192-198.

Magna (2017), "US Advertising Forecast Update March 2017," *IPG Mediabrands.*

Ottaviani, M. (2003), "Overture and Google: Internet Pay-Per-Click (PPC) Advertising Auctions," London Business School - Case Study-03-022.

Varian, H. (2007), "Position auctions," *International Journal of Industrial Organization,* 25(6): 1163–1178.

Varian, H., and C. Harris (2014). "The VCG Auction in Theory and Practice" *American Economic Review: Papers & Proceedings,* 104(5): 442-445.

Vickrey, W. (1961), "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance,* 16(1): 8-37.

Wired (2015), https://www.wired.com/2015/09/facebook-doesnt-make-much-money-couldon-purpose/

Wu, T. (2016), *The attention merchants: the epic scramble to get inside our heads,* Knopf, New York.

CONSUMER SEARCH IN DIGITAL MARKETS

José L. MORAGA GONZÁLEZ[1]

Abstract

The way consumers search in digital markets is different from the way standard models of consumer search presuppose. Specifically, consumer search in digital markets is predominantly *directed*. It is directed because products and/or firms are often heterogeneous so consumer preferences do have a bearing not only on what consumers end up buying but also on the way consumers actively search through the available alternatives. More importantly, firms can affect the direction of search by changing variables that are important to consumers, notably prices. Seeking to properly understand the functioning of digital markets, the consumer search literature has recently been revamped to accommodate search that is directed. It is this new strand of the consumer search literature that I attempt to review in this chapter.

Key words: Digital markets, consumer preferences, prices, consumer search.

JEL classification: D11, L80.

[1] I am grateful to Juanjo Ganuza, Vaiva Petrikaitė, Zsolt Sándor and Matthijs Wildenbeest for their useful remarks.

I. INTRODUCTION

The traditional consumer search literature dates back at least to the seminal article of Stigler (1961), who was surprised by how ubiquitous price dispersion was even in markets for seemingly homogeneous products. Stigler gave the examples of cars and coal and provided price distributions for a particular Chevrolet model and anthracite coal delivered to Washington.[2] He admitted that some of the price dispersion could be due to seller heterogeneity but by no means he accepted that all of it would be. He then conjectured that price dispersion was a firm's response to consumer ignorance in the market, although he did not provide a complete theory where consumer search and price dispersion endogenously arise in market equilibrium. He instead focused on formulating the first steps towards a theory of consumer search.[3]

Stigler's seminal article spawned a great deal of theoretical work that focused on characterizing optimal consumer search. Some years later, probably fueled by Rothschild's (1973) criticism on Stigler's partial-partial equilibrium approach, a significant amount of work appeared centered on finding rationales for equilibrium price dispersion. In Section II, I briefly discuss these two lines of work.

This chapter has a focus on consumer search in digital markets. The way consumers search in digital markets is different from the way traditional models of consumer search assumed. The main difference is that consumer search in digital markets is predominantly *directed*. It is directed because products and/or firms are often heterogeneous so consumer preferences do have a bearing not only on what consumers end up buying but also on the way consumers actively search through the available alternatives. More importantly, firms can affect the direction of search by changing variables that are important to consumers, notably prices. Seeking to properly understand the functioning of digital markets, the consumer search literature has been revamped in recent

[2] Other important works documenting price dispersion are Pratt, Wise, and Zeckhauser (1979) and Lach (2002) and, particularly in digital markets, Baye, Morgan and Scholten (2004).

[3] Stigler presented a model with *simultaneous* consumer search. Simultaneous search, also called *non-sequential* search, refers to a search protocol under which a consumer commits to search at a particular set of firms; once the consumer has visited the firms and inspected their products and/or prices, he/she chooses a firm to buy from, if any. Simultaneous search is in contrast to *sequential* search, which refers to a procedure under which a consumer first searches at a chosen firm and then, upon observing the details of the offer at that firm, the consumer decides whether to buy the product of the firm, continue searching at another firm or quit the market altogether. Morgan and Manning (1985) show that one search protocol is not superior to the other under all circumstances; they argue that often the optimal search rule combines the cost efficiency of simultaneous search with the informational advantages of sequential search.

years to accommodate search that is directed. I review this new strand of the consumer search literature in Section III.[4]

There is still work to do. I discuss recent applications as well as avenues for further research in Section IV. My hope is that the recent theoretical and empirical contributions will inspire additional work.

II. EARLY WORKS

Among the early works characterizing optimal consumer search, it is worth mentioning the contributions of McCall (1970), Kohn and Shavell (1974) and Weitzman (1979). It is important to state right away that, even though these authors focused on consumer search behaviour and assumed the supply side of the market as exogenous, their papers remain very instrumental because any attempt to construct a complete theory of a search market must incorporate optimal consumer search behaviour as a building block.

McCall (1970) (see also Mortensen, 1970) studied the problem of a consumer[5] interested in the purchase of an item who *sequentially* searches for a satisfactory price with infinite horizon. He demonstrated that the optimal search policy is myopic and has the reservation price property. Specifically, the optimal search policy consists of rejecting all prices above a time-invariant threshold price, called the *reservation price,* and accepting any price below it. Moreover, McCall showed that the notion of reservation price, which captures the complex trade-off between the dynamic gains from search and the costs of search, can easily be calculated by myopically equating the gains from searching one more time to the search cost. As expected, the costlier it is to search, the higher will be the reservation price so consumers will be willing to accept higher prices when the search cost goes up.

Because the threshold price is time-invariant, with infinite horizon (or, equivalently, with an infinite number of alternatives) if a price is rejected one time, it will be rejected forever. This implies that with infinite horizon there is no essential difference between search with (costless) recall and search without recall. In later work, Kohn and Shavell (1974) demonstrated that the solution of the sequential search problem with a *finite* number of alternatives,

[4] At this stage it is pertinent to note that there exists another important and large literature in labor economics where search is also directed. However, in that literature workers typically search for availability of a job, not for prices or for product characteristics. The search problem arises because firms are often capacity constrained and, because workers do not coordinate themselves when they apply for jobs, they are uncertain about whether they will get a job. I will not review this literature here. For a recent survey of this literature, see Wright, Kircher, Julien and Guerrieri (2017).

[5] In his original work McCall treats the case of a worker searching for a well-paid job but that problem is isomorphic to that of a consumer searching for a reasonably priced item.

arguably more realistic in many real-world settings, also has the time-invariant reservation value property when recall is costless. When recall is costly, by contrast, the reservation value decreases as search proceeds (see *e.g.,* Janssen and Parakhonyak, 2014).

Weitzman (1979) extended the theory of consumer search in a crucial direction. He did allow for ex-ante heterogeneity in the alternatives available. Specifically, he considered the problem of a decision maker who faces a finite number of different options of unknown value. In such cases, the solution to the search problem not only consists of a stopping rule but also an order of search. Weitzman demonstrated that the optimal search policy consists of ranking the options in terms of reservation utilities (a notion similar to that of reservation price), searching them in declining reservation utility order and stopping search when the highest observed utility is greater than the reservation utility of the next option to be searched.

Chade and Smith (2006) is also a critical step towards a better understanding of directed consumer search. They studied the same search environment as in Weitzman (1979) but modelled simultaneous search rather than sequential search. This means that consumers have to optimally choose a subset of ranked options to maximize expected utility. They showed that the problem is very hard in general but provided an algorithm that can solve the problem when the consumer payoff satisfies a regularity condition or the utility distributions of the different alternatives can be ranked according to second-order stochastic dominance. The algorithm, called *marginal improvement algorithm,* is simple. Options are first ranked according to expected utility. The option with the highest expected utility is added to the optimal set provided that the expected utility is greater than the cost of inspecting the option. The second option is added to the set provided that the expected maximum utility of the two options in the set minus the expected utility of the first option, that is the marginal increase in expected utility, is greater than the search cost. And so on and so forth.

Somewhat surprisingly, Weitzman's and Chade and Smith's characterizations of optimal consumer search among heterogeneous alternatives have received little attention until recently, perhaps because of the difficulties to model a supply side with heterogeneous firms. As it will become clearer later in Section III, their contributions constitute a critical stepping-stone to the development of the theory and the empirics of markets with directed consumer search.

As mentioned above, another relevant line of work focused on finding rationales for equilibrium price dispersion. Stigler (1961) argued that price dispersion was probably the outcome of search frictions in the market but did not supply a model that featured price dispersion and search in market equilibrium. Diamond (1971) went further and demonstrated that with

sequential consumer search it is not possible to have price dispersion and search at the same time. His famous result, which is known as the *Diamond-paradox,* is better understood within the context of a simple example. Suppose that there is a market in which N homogeneous product sellers compete in prices to sell their goods to identical consumers. Suppose consumers have a valuation for the product given by $v > 0$, search sequentially to find a reasonable price, and incur a common positive search cost $c < v$ each time (beyond the first one) they search. Diamond (1971) showed that there cannot be an equilibrium with price dispersion and, as a consequence, consumers will not search; moreover, the unique equilibrium must have all firms charging the monopoly price, which, in this case, equals the willingness to pay of consumers, v. As Stiglitz (1989) put it, if an equilibrium with price dispersion existed, then the firm charging the lowest price in the market would regret it and immediately deviate by raising the price by an amount not exceeding the search cost of consumers, c. Such a deviation would increase the margin of the deviant firm without reducing its demand, which would result in higher profits.[6]

Rob (1985) and Stahl (1996) demonstrated that the Diamond result is quite robust in markets where products are homogeneous and consumers search sequentially, even if consumer search costs are heterogeneous. Stahl (1989) (see also Stahl, 1996) argued that there is one particular case of consumer search cost heterogeneity that can cause price dispersion and search to emerge jointly in equilibrium. This happens when a fraction of the consumers has zero search costs.

In Stahl's (1989) contribution, a finite number of symmetric firms selling a homogeneous product compete in prices. Some consumers have a common positive search cost and search sequentially to find a reasonable price; the rest of the consumers have no search costs at all and buy from the firm that charges the lowest price in the market. It is easy to understand that in any equilibrium prices must be dispersed. To see this, notice first that consumers with positive search cost will optimally adopt a stopping rule characterized by a time-invariant reservation price. Consumers with zero search cost will buy from the firm offering the lowest price in the market. Now, suppose that all the firms charged the same price in equilibrium. If this were so, an individual firm would have an incentive to slightly undercut that price. The reason is that such undercutting would attract all the consumers with zero search cost without compromising

[6] More concretely, suppose that there is an equilibrium with price dispersion; because firms are symmetric, this means that firms' profits must be the same no matter the price they charge. Because consumers do not see prices before search, it is reasonable to expect that, in their first search, they will randomly choose one of the firms and pay it a visit to inspect its price. Consider now a firm charging the minimum price in the support of the price distribution. This firm could raise its price by just less than the search cost c and its (randomly allocated) consumers would not leave the store to conduct a second search for a better deal. This deviation would then give the firm higher profits. This simple argument is rather powerful and rules out asymmetric equilibria with different prices as well as symmetric equilibria with prices less than v. The only equilibrium candidate left is the monopoly price equilibrium.

the per consumer margin, which would result in higher profits. This reasoning rules out a single price equilibrium in which the price is higher than the marginal cost. But, the argument continues, marginal cost pricing is not an equilibrium either because, by the logic behind the Diamond paradox explained above, an individual firm could raise the price without compromising its sales to the consumers with positive search costs and increasing its margin. In conclusion, the tension between charging low prices to attract the consumers with zero search cost and charging high prices to take advantage of the consumers with positive search costs is balanced when firms randomize their prices. The next step in the analysis is to understand that in any equilibrium, the price distribution has to be continuous, that the support has to be convex and that the maximum price has to be equal to the reservation price of the consumers with positive search cost (for details I refer the reader to the original contribution of Stahl). The last observation implies that no consumer will search beyond the first firm.[7]

Stahl's model is one of the most celebrated search models because it offers a richness of results within one common and relatively tractable setting. The mixed pricing equilibrium moves continuously from the Diamond paradox (monopoly pricing) to the Bertrand paradox (marginal cost pricing) as we increase the share of consumers with zero search cost from 0 to 1. Moreover, the distribution of prices becomes higher (in the sense of first order stochastic dominance) as the search cost increases. Finally, an increase in the number of competitors in the market results in higher prices on average.

Stahl probably got inspired by the famous *model of sales* of Varian (1980). In Varian's model there are some consumers who buy from the firm charging the lowest price in the market, while the remaining consumers buy from one of the remaining firms chosen at random. In equilibrium, by the same logic as in Stahl (1989), there is price dispersion. Varian argued that his informational assumption on the demand side of the market could easily be made endogenous assuming consumer search is all-or-nothing in the sense that consumers who pay the search cost learn the prices of all the firms in the market. A fine application of the model of Varian (1980) to search engines is Baye and Morgan (2001).

Another key contribution towards a better understanding of the joint occurrence of equilibrium price dispersion and search is Burdett and Judd (1983). Burdett and Judd consider a market in which infinitely many firms compete in prices to sell a homogeneous item to symmetric consumers who search non-sequentially for lower prices. They show that, in addition to the Diamond

[7] Strictly speaking, Stahl (1989) is thus a model of search without proper search in the sense that consumers with positive search cost do not search beyond the first firm. Notice, however, that this is due to the assumption that all consumers with positive search cost have the same search cost. If they did have different search costs, they would have different reservation prices and there would be active search in equilibrium (see Stahl, 1996). For an application, see Giulietti, Waterson and Wildenbeest (2014).

equilibrium, there exists an equilibrium featuring search and price dispersion. The market equilibrium has the following characteristics. Consumers, correctly expecting prices to be dispersed, randomize between searching one time and searching two times.[8] Firms, correctly expecting consumers to mix between one and two searches, by the same logic as in Varian's and Stahl's papers again, mix their prices in equilibrium.

Burdett and Judd (1983) is a very influential paper because, to my knowledge, it is the first paper to obtain price dispersion without any ex-ante heterogeneity in the market whatsoever. Janssen and Moraga-González (2004) extended their setting to the case of oligopoly and allowed for some consumers to have zero search costs. They showed that the average price is non-monotonic with respect to the number of competitors in the market, first decreasing and then increasing.

The previously discussed literature features models in which consumers search for prices. In real-world markets, whether digital or not, it is common for consumers to visit shops to find out about additional product characteristics. Wolinsky (1986) is an inspiring paper where firms sell differentiated products and consumers search sequentially to find a product that is satisfactory. Wolinsky demonstrates that a pure-strategy equilibrium exists under quite reasonable conditions. The equilibrium price is below the monopoly price so the Diamond result does not hold in this setting with differentiated products. The reason is that when products are differentiated consumers will search even if all prices are equal in equilibrium. Consumer search for a good match disciplines the firms, which end up charging prices below monopoly. Another interesting observation is that the equilibrium price is above the marginal cost even under the assumption of free entry of firms. Search costs thus constitute a source of market power and thereby competitive markets with search costs can be regarded as a foundation for monopolistic competition.

Though nowadays Wolinsky's paper is regarded as the work-horse model of consumer search for differentiated products, it took quite a few years till Wolinsky's work saw applications in Industrial Organization. Anderson and Renault (1999) developed further Wolinsky's framework. They showed that

[8] Notice than when consumers are similar there is no equilibrium in which consumers search two times or more. This is because in such a case, by the Bertrand logic, all the firms would charge a price equal to the marginal cost; but if all prices are equal, there is no point in searching that much. With search cost heterogeneity, some consumers will search once while others twice, thrice etc., a point first made by Hong and Shum (2006) and further elaborated by Moraga-González, Sándor and Wildenbeest (2017a). The latter paper provides evidence that the relationship between prices and consumer surplus with respect to the number of competitors depends upon the nature of search cost dispersion. When search costs are very dispersed, the average price increases while consumer surplus may decrease in the number of competitors. When search costs are little dispersed, the average price decreases and consumer surplus increases in the number of active firms in the market.

the equilibrium price increases in search cost, a result that is quite obvious in the case of free entry of firms but more intricate to show in oligopoly. Anderson and Renault also studied how the equilibrium price depends on the extent of product heterogeneity and the number of competitors. They found that the equilibrium price can decrease in product differentiation. This is somewhat surprising because price tends to fall as products are more heterogeneous; however, more product heterogeneity increases the incentives to search and this effect can be stronger. Anderson and Renault also studied the effect of firm entry on the equilibrium price. They found that the equilibrium price decreases in the number of firms and, moreover, that the market with search frictions tends to have too many firms compared to the socially optimal number of firms.

Moraga-González, Sándor and Wildenbeest (2017b) generalize the Wolinsky' framework to the case in which consumers have heterogeneous search costs. They argue that an unsatisfactory feature of most of the search models is that they assume that all consumers search. This sort of "fully-covered-market" assumption is somewhat at odds with the idea that consumers vary in their costs of search. As a matter of fact, unless one is prepared to assume that consumer search cost heterogeneity is limited, there must be consumers out there who do not find it worthwhile to search for a particular product. Admitting this implies that an increase in search costs has a bearing on two margins. First, an increase in search costs affects negatively the intensive search margin, or search intensity. By this effect, demand tends to become more inelastic and prices tend to increase. Second, an increase in search costs affects negatively the extensive search margin in the sense that more consumers will decide to not search at all. If consumers did not adapt their search intensity, by this effect demand would become more elastic and prices would tend to increase. Moraga-González, Sándor and Wildenbeest provide conditions on search cost densities under which one effect dominates the other and viceversa.[9]

III. DIRECTED CONSUMER SEARCH

The influential work presented in Section II refers to markets where consumers search randomly. That a firm is visited by consumers is thus merely driven by the bare existence of the firm, not by its attractive price, high quality or good location. In digital markets, but also in some conventional markets, things are quite different. First, it is very easy for consumers to compare prices; as a matter of fact, consumers sometimes sort alternatives on the basis of prices, and proceed by inspecting first the options priced more attractively

[9] Moraga-González, Sándor and Wildenbeest (2017b) results can help understand Hortaçsu and Syverson's (2004) empirical observation that prices went up in the US mutual fund industry during 1990's despite the observed decrease in search costs.

before moving to the more expensive ones. Second, it is quite natural that other product characteristics, not only the price, are readily observable. For example, a consumer who uses an online travel agent to book a hotel often sees the name of the hotel, number of stars, location, review score, a photography, etc. without incurring much search effort. The information easily made accessible reveals a great amount of product heterogeneity and it is precisely the interaction between product heterogeneity and consumer tastes that guides consumer search. It is reasonable to expect, therefore, that consumer search in digital markets is *directed,* and that, reflecting the heterogeneity of product features and consumer preferences, the distribution of consumer visits across firms is quite unequal.

De los Santos (2018) presents evidence on consumer search patterns for books using 2002 and 2004 data from the ComScore Web-Behavior Panel. The average buyer, who bought 2.2 books in 2002 and 2.4 books in 2004, visited 1.2 stores in 2002, and 1.3 in 2004. Only around 30% of the consumers searched at more than one firm. In only 25% of the book purchases had consumers searched at more than one firm. This is evidence of there being relatively little search, but is also consistent with the idea that if consumers search for a good price of a specific book, then much search is likely to be suboptimal (cf. Burdett and Judd, 1983). De los Santos also points out that the distribution of searches is quite unequally divided across firms, with a strong bias towards the major book sellers Amazon and Barnes & Noble. Specifically, buyers visited Amazon in 74% of the book purchases while only 17% of the buyers from Amazon visited other bookstores. Of the buyers of Barnes & Noble, 39% visited at least another bookstore. Regarding the order of search, Amazon was searched first in 65% of the sample, while Barnes & Noble in only 17%. Among those who bought a book from Amazon, 91% visited first Amazon, while among those who purchased from Barnes & Noble 68% visited first Barnes & Noble.[10]

Arbatskaya (2007) is one of the earliest papers in which consumers do not search randomly. Because consumers do not decide the order in which they search, it is more accurate to regard her paper as one where search is ordered, but not directed. Firms sell homogeneous products, compete in prices and consumers search sequentially in a pre-specified order known to the firms to find an attractive price. Consumers have heterogeneous search costs. Arbatskaya shows that prices must decrease in search order. The intuition behind the pricing result stems from the observation that only consumers with

[10] See also De los Santos, Hortaçsu and Wildenbeest (2012), who use a similar dataset to test among the theories of sequential and simultaneous consumer search. They conclude that simultaneous search is more in line with what consumers actually do when they search on the Internet. See also Honka and Chintagunta (2017), who, using data on consumer search and purchase for auto insurance in the U.S., provide support for the simultaneous search protocol.

relatively low search costs are prepared to walk away from the firms that appear early in the search order to venture firms that come later. Firms that appear later in the search order, knowing that only more elastic consumers patronize their shops, have an incentive to charge a lower price. The equilibrium exhibits price dispersion and active search.

When products are differentiated as in Wolinsky (1986), the logic of Arbatskaya need not work because consumers may decide to leave firms that appear early in the search order in an attempt to find better products even if they expect higher prices later on. To model this idea, Armstrong, Vickers, and Zhou (2009) consider a market where consumers visit one prominent firm first and, if its product is not satisfactory, they continue searching, in this case randomly, among the remaining firms. Armstrong, Vickers, and Zhou show that the prominent firm charges a price lower than the price charged by the non-prominent firms, provided that the search cost is strictly positive.[11] This result originates from the observation that only consumers who are disappointed with the product offered by the prominent firm end up visiting the non-prominent firms to inspect their products. Knowing this, the non-prominent firms have incentives to increase their prices relative to the prominent one because they face a pool of consumers who are more inelastic. Despite charging a lower price, the prominent firm makes higher profits than the non-prominent firms,[12] providing a theoretical foundation of the "proverb" that *being first is best*.[13]

[11] When the search cost converges to zero, every consumer visits every firm before picking a product and, consequently, firms end up charging essentially the same price. Thus, prominence loses its value. In Rhodes (2011), by contrast, prominence has value even if search is costless. The key difference is that Rhodes assumes that consumers know the valuations they place on the products offered by the firms but ignore which firm sells which product. Consumers learn which product is sold by a particular firm after paying it a visit. Even if search is costless, there is no reason for a consumer to continue searching after she has found the best match. Because all consumers search first the prominent firm, a non-prominent seller thus knows that it attracts consumers who place a high value to its product, and therefore it charges a high price. The prominent firm charges a lower price but has a larger demand and earns higher profits, even if searching is virtually costless.

[12] Ursu (forthcoming) presents empirical evidence from consumers searching for hotels in the online travel agent Expedia that the position of a product in the Expedia list has a causal effect on clicks, but conditional on clicking, it does not affect the likelihood of a purchase.

[13] Fishman and Lubensky (2018) modify Armstrong, Vickers and Zhou's framework by explicitly accounting for return costs. When consumers have both costs of search and costs of returning to previously visited sellers, a trade-off arises when considering the incentives to be first in the search order of consumers. Being searched first is advantageous if the consumers find good values at the firm because then search and return costs lower the incentives consumers have to search further. But being second is advantageous when consumers are likely to find bad values at the first firm because then the return costs will prove pivotal to make consumers "stay" with the second firm even if the first firms turns out to be better *a posteriori.* Fishman and Lubensky show that for increasing utility densities first is better while for decreasing utility densities second is better. With N firms, any position can be best depending on the utility distribution, but an increase in the number of firms makes the first position more favourable relative to any other, which is in line with some recent empirical results (see *e.g.* De los Santos and Koulayev, 2013).

Making a firm prominent typically leads to higher industry profits, at the expense of consumer surplus. Total welfare also decreases when one firm is prominent.

Zhou (2011) extends the paper of Armstrong, Vickers, and Zhou by considering a situation in which consumers search sequentially through N options in a pre-specified order known to the firms, like in Arbatskaya (2007). He shows that prices must increase in the order of search, thus generalizing the Armstrong, Vickers, and Zhou result. Zhou's equilibrium also exhibits price dispersion and active consumer search but, in contrast to Arbatskaya's result, without the need of search cost heterogeneity. Zhou also shows that, compared to random search, ordered search may result in overall higher prices when there are enough firms in the market.

1. Influencing the Direction of Search

The papers described above assumed that consumers check the prices of the firms or inspect the suitability of the available options in an exogenously specified way. I move now to discuss research in which consumers choose the order in which they inspect the various alternatives and firms can take actions to influence this order.

Wilson (2010) is one of the first papers modelling the idea that firms can affect the ease with which consumers can find their deals, thereby influencing the order of search.[14] Wilson considers a duopoly model similar to Stahl (1989) with shoppers and non-shoppers and allows the firms, prior to competing in the market, to pick the search cost of consumers. He demonstrates that an equilibrium where both firms pick zero search cost does not exist. If such an equilibrium existed, firms would make zero profits. In that situation, an individual firm would gain by deviating by raising the search cost. Though the deviant firm would decrease its appeal for the non-shoppers and would decrease its volume of sales, this deviation would relax competition for the shoppers and increase the profit margin. Wilson (2010) shows how starting from a symmetric situation, the market forces can lead to an asymmetry in the cost consumers have to incur to visit the firms. This reduces consumer welfare. His results can potentially explain why not all firms choose to go online, where search costs are arguably lower.

[14] Ellison and Wolitzky (2012) is another paper modelling the idea that search costs are endogenous but because consumers do not observe the search cost of the firms before they visit, firms cannot affect the order of search, not even off the equilibrium path. They use the Stahl (1989) setting but introduce the idea of diseconomies of search in the sense that search costs increase convexly (rather than linearly as usual) in the number of visits. Ellison and Wolitzky find a symmetric equilibrium where firms pick the highest possible search cost, thereby weakening competitive pressure and raising profits.

Haan and Moraga-González (2011) is another early attempt to model situations where firms can influence the order of consumer search. They do so in the framework of Wolinsky (1986) and Anderson and Renault (1999) where symmetric firms sell differentiated products and consumers search sequentially to find a satisfactory good. Because of reasons that will become clearer later, Haan and Moraga-González consider a situation in which firms compete for the attention of consumers via advertising, and not via lower prices. In their model, a firm that advertises better or more, which is assumed to be significantly costlier, becomes more salient in the marketplace and therefore attracts a higher share of the consumers who, at any given moment, contemplate conducting another search. Although advertising does not alter consumers' willingness to pay, consumers increase the propensity with which they buy the product of a firm when they see that this firm advertises more than the rest.[15] Firms find themselves in a classic prisoners' dilemma. If a firm advertised less than its rivals, it would probably be relegated to later positions in the search order of consumers, or even to the very end of it. In equilibrium, all firms advertise with the same intensity to gain consumer attention and advertising is purely wasteful. Haan and Moraga-González show that in equilibrium prices increase in search costs. This price increase raises the reward a firm obtains when winning the race for consumer attention and, consequently, results in greater incentives to advertise. Together, these two effects may cause profits to decrease as search cost goes up.

One of the obvious ways in which firms can favorably affect the order of consumer search is by quoting lower prices. However, the modelling of price-directed search has proven quite difficult and only very recently there has been enough advancement. As discussed in Armstrong and Zhou (2011) and Haan, Moraga-González, and Petrikaitė (2017), if prices were observable prior to search in the standard model of Wolinsky (1986) and Anderson and Renault (1999), a pure-strategy symmetric equilibrium would fail to exist. The reasoning is as follows. Suppose all firms charged a price strictly higher than the marginal cost in symmetric equilibrium. In that case, consumer search would be random and no firm would be visited first, second, third etc. with a probability different from the other firms. If a firm deviated by slightly undercutting the equilibrium price, then all consumers would start their search at that firm, which would lead to a discontinuous increase in its demand without compromising its margin. Such a deviation would thus be profitable. This logic suggests that only marginal cost pricing could be a pure-symmetric equilibrium. However, marginal cost pricing is not an equilibrium either because an individual firm would find it profitable to deviate to a higher price. Despite the fact that this firm would be relegated

[15] Advertising is thus *persuasive* in Haan and Moraga-González (2011) but, in contrast to the traditional notion of persuasive advertising in the economics literature, willingness to pay is not affected so a sound welfare analysis can be conducted.

to the very end of the consumer search order, still those consumers unsatisfied with the offerings of the rival firms would visit it and end up buying there. Such a deviation would thus be profitable. The failure of existence of a pure-strategy symmetric equilibrium need not be a problem in itself, but what happens is that the characterisation of the mixed-strategy equilibrium has proven to be non-tractable.

The recent literature has overcome this difficulty in two ways. One approach, exemplified by Armstrong and Zhou (2011) and Ding and Zhang (2018), has consisted of modifying the model in order to obtain enough tractability to compute the mixed-strategy equilibrium. The other direction, illustrated in Haan, Moraga-González, and Petrikaitė (2017) and Choi, Dai and Kim (2017), has involved enriching the model of product differentiation to restore the existence of a pure-strategy equilibrium. I now discuss these two approaches in some more detail.

Ding and Zhang (2018) is one of the earliest papers with price-directed search.[16] Ding and Zhang, aiming at modelling consumer search in situations in which prices are readily observable by consumers, introduce a simple form of product differentiation into the seminal paper of Stahl (1989). Specifically, they assume that firms carry products that may or may not fit the tastes of consumers; moreover, whether their products fit or do not fit is random across consumers and firms. This simple form of product differentiation is a smart device to allow for search being directed by prices, while still keeping the model tractable. In their model, like in Stahl (1989), there are shoppers and non-shoppers. Shoppers know which products meet their needs and the prices at which they are sold so they pick the cheapest of the products matching their needs, if there is any. Non-shoppers search through the firms sequentially and in order of increasing prices with the same aim, that is, in order to check whether there are products that suit them. A nice feature of Ding and Zhang's model is that it collapses to Stahl (1989) when products fit with certainty and prices are not observable before search.

There are a few results in the paper of Ding and Zhang (2018) worth highlighting. A first interesting result pertains to the way consumers search. They show that non-shoppers will never search at firms charging a price higher than a threshold price. Such a threshold price happens to increase as search costs decrease, reflecting the fact that consumers are prepared to search at higher price firms if their search cost becomes lower.

The second result is that an equilibrium in mixed strategies exists and it can be characterized explicitly. Price dispersion arises for reasons similar to Stahl (1989). Namely, because the probability a product matches the tastes of

[16] To the best of my knowledge, the earliest version of this paper is by Zhang alone and dates back to 2011.

a consumer is less than one, there may be shoppers and non-shoppers among the matched buyers of a firm. Because it is likely that the matched shoppers also match with other firms, the firm has an incentive to charge low prices. At the same time, because non-shoppers have to pay search costs to check if they match with other products, it is less likely that they will do so and therefore firms also have an incentive to charge high prices. Like in Stahl's paper, these two incentives are balanced when firms randomize their prices. Interestingly, the mixed strategy equilibrium may have a non-convex support; specifically, when the search cost is high, the firms draw their prices from two disjoint sets of prices. This happens because when the search cost is high, the threshold price above which non-shoppers decide to not visit a firm is sufficiently low.[17] When the search cost decreases, firms optimally increase the probability of charging a price from the low-price interval. However, Ding and Zhang show that, surprisingly for a search model, the average price in the market can increase as the search cost falls. The reason for this is that a lower search cost increases the maximum price firms can offer to non-shoppers to entice them to visit the firms in order to inspect their products.

Armstrong and Zhou (2011) introduce search frictions in a duopoly market where firms sell products that are differentiated à la Hotelling. Prices can be easily accessible via a website, in which case consumers, who still need to check the suitability of the products, will inspect first the product of the firm that charges a lower price. Because of the special structure of the Hotelling preferences, a consumer only needs to make one search in order to discover the value she places on both products. By the logic mentioned above within the context of price-directed search, it is easy to see that there is no pure-strategy equilibrium in prices. However, the Hotelling preferences allow for the explicit characterization of the mixed strategy equilibrium, which features a continuous density function. In equilibrium prices decrease as search cost goes up, which, as mentioned above, is in contrast to most search models. In this case, what happens is that when consumers' search costs go up, they become more unwilling to search beyond the first firm. This makes being first in the search order of consumers more valuable, which gives firms a stronger incentive to compete for that position.

Haan, Moraga-González and Petrikaitė (2017), to my knowledge, were the first[18] to propose building additional product differentiation into the Wolinsky framework in order to restore the existence of a price equilibrium in pure strategies. In their duopoly model, firms sell products with two attributes, both of them horizontally differentiated. The key assumption is that

[17] Interpreting prices from the low interval as sales prices and prices from the high interval as regular prices, this equilibrium is consistent with empirical evidence on pricing by e-retailers on the Internet (Baye, Morgan, and Scholten, 2004).

[18] The first version of this paper is by Haan and Moraga-González and dates back to 2011.

one attribute is observable before search and the other only after search. The first attribute thus represents product characteristics that can easily be observed for example in a website. The second attribute represents search characteristics, that is, properties of the product that can only be ascertained upon close and careful inspection. In many online situations consumers confront this search problem. For example, when looking for a flight to a particular city destination some product characteristics are often readily observable like the name of the airline, destination airport, price, and flying times. However, other characteristics of the service such as terminal of arrival, air-miles bonuses, meals, luggage policy, administration fees, etc. are only observable upon careful reading of the flight details.

In this model, because some product characteristics are observable before search, search is already naturally directed. In addition, firms can favorably affect the direction of search by quoting lower prices. As intuition would suggest, provided that there is sufficient differentiation in the product attributes that are readily observable, an equilibrium in pure strategies exists. The logic mentioned before that a firm that slightly undercuts the equilibrium price sees its demand jump up does not apply here because consumers not only care about the price when they choose where to start searching for a satisfactory product. The price might be sufficiently low but if the other observable characteristics are not good enough, it will be very difficult to entice a consumer to visit. Haan, Moraga-González and Petrikaitė compare the price equilibrium when the price is readily observable with that when the price is not, as in the standard model of Wolinsky (1986) and Anderson and Renault (1999). They find that the equilibrium price is always lower when consumers observe the prices of the firms before starting search. The reason is that, when prices are observable before consumers start searching, a cut in the price not only increases the chance that consumers stop searching at that firm but also increases the chance they visit it. When the price is observable before search, the demand of a firm is thus more elastic and therefore prices are lower in equilibrium. Haan, Moraga-González and Petrikaitė also study the comparative statics effects of higher search costs. They show that when firm prices are observable before search, they decrease as search costs increase. The intuition is similar to that in Armstrong and Zhou (2011). When search costs go up, consumers are less likely to walk away from the firm they visit first, which gives firms stronger incentives to compete in the contest for being first. Interestingly, despite troubling consumers, higher search costs may be good for them due to this lowering price effect.[19]

[19] When the direction of search is influenced by prices, a direct link is established between the price and the propensity consumers have to visit the firm. It is this link that produces the unconventional result that higher search costs lead to lower price and profits. In Garcia and Shelegia's (forthcoming) paper on observational learning, the price is not observed but nevertheless it has a bearing on the number of consumers that visit the firm in the future. Because of this, they also find that equilibrium prices may decrease as search costs increase.

Haan, Moraga-González and Petrikaitė (2017) performed their analysis within the context of a duopoly model. Extending the analysis to oligopoly proved to be a difficult challenge. The problem is that they compute demand by explicitly taking into account the different search paths consumers may follow before they buy from a given firm. With just two firms, the demand of a given firm stems from three groups of consumers. Specifically, one group is comprised of consumers who start searching at the firm in question and stop after finding a suitable product; the second group is made of consumers who start searching at the rival firm, do not find something satisfactory there and move to the firm in question where they do find something they like; and finally, consumers who start searching at the given firm, go to check the product of the rival firm but decide to return to the former to buy there. With three firms, there are eleven different search paths a consumer can follow before purchasing from a specific seller. As the number of sellers grows, the number of search paths increases factorially.

Armstrong (2017) and Choi, Dai, and Kim (forthcoming) have independently solved the problem of computation of demand in general settings. They show that to compute demand one can dispense with the myriad of search paths consumers can follow and reformulate the problem as a static discrete-choice problem in which consumers choose the alternative that gives them the highest minimum of the reservation utility and the realized utility among all available alternatives.[20] Intuitively, the reason why the minimum of the reservation utility and the realized utility is what matters for a purchase has to do with the fact that both have to be relatively high. In fact, before an option is bought, it must be searched, in which case the reservation value should be relatively high. Moreover, the realized utility must also be relatively high for otherwise consumers would not buy the current alternative and continue searching.

The reformulation is as if the search paths consumers can follow before they buy the product of a firm get "integrated out" and thereby demand has a relatively simple and well-known expression. More importantly, what is known from discrete-choice models applies to the sequential consumer search model and for example the existence and uniqueness of equilibrium can be established invoking results from that literature. Choi, Dai, and Kim further show how to perform comparative statics analysis using the distribution of the minimum of the reservation utility and the realized utility. They show that, in the absence of an outside option, the equilibrium price will increase as the distribution of the minimum of the reservation utility and the realized utility becomes more dispersed. An increase in product differentiation typically does so and therefore results in higher prices. This outcome is in contrast with the result mentioned above in Anderson and Renault (1999). In their paper the equilibrium price

[20] In a sense, this possibility was anticipated by Armstrong and Vickers (2015) who noted that, under some assumptions, the sequential search model produces demands that are consistent with discrete choice.

decreases as product differentiation goes up because consumers search more. The main difference is that in Anderson and Renault prices are not observable before search. Choi, Dai, and Kim also show that an increase in search costs lowers the dispersion of the minimum values and therefore the equilibrium price decreases, so the result obtained by Haan, Moraga-González and Petrikaitė (2017) for duopoly holds more generally.

Particularly in digital markets where platforms have become central market places, another way in which firms can affect the order of search is by bidding payments to platforms to be placed high on the list of search outcomes associated with a given search query.[21] Athey and Ellison (2011) and Chen and He (2011) are the first papers presenting models where horizontally differentiated firms bid for placement in lists of search results and consumers search sequentially through the listed options.[22]

In Chen and He (2011), firms are heterogeneous in regard to the probability with which they can satisfy consumer needs. Consumers know such probabilities but they do not know which firm has which probability of being suitable, which simplifies the analysis. Consumers decide how to search through the list of options presented to them. Chen and He show that a separating equilibrium exists in which more suitable sellers bid higher payments than less suitable ones, whereby the order in which the options are presented reveals the quality of the firms. Correspondingly, consumers optimally search from top to bottom. The separating equilibrium has more efficient search, higher output and social welfare than when consumers search randomly. Athey and Ellison (2011) present a more general incomplete information structure where the suitability probabilities are random draws from a distribution. This makes the analysis substantially more complex because consumers have to update their beliefs about the suitability probabilities as they search. They nicely characterize an equilibrium similar to that in Chen and He. In both these papers, the pricing of the alternatives listed does not play much of a role.

Chen and He (2011) and Athey and Ellison (2011) model the interaction between consumer search and firm bidding for positions in settings where the pricing of products does not play a significant role. Specifically, in Athey and Ellison (2011) the pricing is exogenous while in Chen and He (2011), conditional on matching the tastes of consumers, products are homogenous so by the logic of Diamond (1971) the unique price equilibrium is the monopoly price.

[21] See also Armstrong and Zhou (2011) for a model in which firms pay commissions to intermediaries to see their products promoted.

[22] The start of this line of work goes back to Varian (2007) and Edelman, Ostrovsky and Schwarz (2007), who studied optimal bidding in position auctions. They did not consider, however, the bidding problem in connection to a search environment.

Anderson and Renault (2017) add to this line of work by presenting a model that incorporates bidding for positions, product pricing and consumer search. To do so, they modify the framework of Wolinsky (1986) and Anderson and Renault (1999) by allowing for firm heterogeneity and use Weitzman (1979) rule to characterize consumer search behaviour. They cleverly modify consumer preferences to avoid that consumers return to previously visited options, which makes the analysis tractable. Anderson and Renault show that equilibrium order of search is linked to pricing, not to bidding, in contrast to Chen and He (2011) and Athey and Ellison (2011). Their most important result is that in their more general model there is a misalignment between the order preferred by the firms and the order preferred by the consumers. This does not occur in the simpler settings of Chen and He (2011) and Athey and Ellison (2011).

IV. CONCLUDING REMARKS

In digital markets, but not only, the order in which consumers search through the available alternatives is dictated by what they know a priori about them, which very often, though not always, includes their prices. This means that consumer search is quite different from the way traditional models of consumer search are constructed. In this chapter, I have started by summarizing key classical contributions to the literature on consumer search, and then continued by explaining recent advances that make the consumer search apparatus more suitable to address theoretical and empirical challenges that help better understand the functioning of digital markets.

I would like to finish this chapter by mentioning areas of work that have developed in parallel and have benefited or could benefit in the future from the recent advances. While doing so, I will also describe some avenues for further research.

The first area worth mentioning is the empirical studies on estimation of demand for differentiated products and the assessment of market power. The standard assumption in this work is that consumers have perfect information about all the products available in the market (see *e.g.,* Berry, 1994; Berry, Levihnson and Pakes, 1995; and Nevo, 2001). This, arguably, is by no means a reasonable assumption in many real-world markets because consumers often ignore, or partially ignore, the utility they get from the various alternatives, either because they do not know the prices at which they sell and/or because they have to carefully inspect the products to discover all the characteristics. Moreover, if consumers have partial information prior to search, their search strategy will naturally be directed. Ignoring that the set of alternatives consumers consider is endogenous is likely to lead to biases in the estimates

of consumer preferences and market power. In order to deal with this problem, new methods are necessary.

Acknowledging that consumer search models of demand are better suited for making inferences in some real-world settings, a crucial issue is the identification of search costs. In environments with differentiated products and heterogeneous preferences, the identification of search costs is challenging.[23] The reason is that the impact of search costs and preferences on choices may be difficult to separate. For instance, if the market share of a firm is relatively low, is it due to low tastes for the products sold by this firm or by high search costs? Likewise, if a consumer is observed to walk or click away from the product of a firm, is this due to the consumer placing a low value for the product of the firm or to the consumers having a low search cost?

Moraga-González, Sándor and Wildenbeest (2017c) propose an empirical approach to estimate demand in the automobile market allowing for directed sequential consumer search. They adapt the Armstrong (2017) and Choi, Dai, and Kim (forthcoming) approach by allowing for search cost heterogeneity and multiproduct firms and estimate demand in the well-known framework of Berry, Levihnson and Pakes (1995). To estimate search costs, they exploit variation in the costs of visiting dealerships. They find that the estimates of search costs are significantly different from zero. The search cost model produces less elastic demands and therefore firms possess greater market power than in the full information model. In future work, it would be useful to investigate the optimality of dealership networks. More broadly, future papers on the theme could allow for more general models of search, for example, by incorporating search for quality and price bargaining.

Ershov (2018) is another paper that takes advantage of variation in search costs. Exploiting a natural experiment in the Google Play mobile apps store that

[23] Using a heterogeneous search costs version of Burdett and Judd's (1983) model of simultaneous search, Hong and Shum (2006) were the first to present a structural methodology to retrieve search costs in markets for homogeneous goods using only price data. Moraga-González and Wildenbeest (2008) extended their approach to the case of oligopoly and presented a way to estimate the search cost distribution by maximum likelihood. Moraga-González, Sándor and Wildenbeest (2013) demonstrated that the search cost distribution cannot be non-parametrically identified in its full support using price data from a single market, even if there are infinitely many firms participating in the market. They showed that combining price data from many product markets where consumers face the same search costs identifies the entire search cost distribution and provided a semi-non-parametric approach to estimate it using this kind of data. Sanches, Silva Junior and Srisuma (forthcoming) propose a minimum distance approach to estimate the search cost distribution. De los Santos (2018) show how to use search data, in addition to price data, to estimate the model allowing for unequal visiting probabilities. Finally, Hortaçsu and Syverson (2004) show that when price and quantity data are available, this methodology can be extended to richer settings where price variation is not only caused by search frictions but also by quality differences across products.

reduced the search costs for game apps and not for other apps, he estimates the effects of lower search costs on entry, product design and quality of the apps. He finds more entry but less quality in the treated group than in the control group. In the future, more work should be dedicated to understand how search costs affect entry and quality investment in search markets, certainly in directed search environments.[24]

In most cases it is difficult to exploit variation in search costs just because such data are rarely available. Internet data are a great advantage because the econometrician not only observes purchases but also search/click behaviour. Koulayev (2014) shows how detailed data on browsing and clicking on the internet can be used to identify search costs that rationalize sequential search. His data comes from a search engine for hotel bookings. After a search query, the buyer observes a page containing a first set of search results. Then, the buyer has to click to proceed to another page of search results, and so on. If the econometrician observes the first set of search results and the posterior clicking behaviour, then changes in the observed products across searchers provide a source of variation that allows for the estimation of search costs. Following this line of reasoning, Kim, Albuquerque and Bronnenberg (2010) exploit search data from the recommendation system of *Amazon.com* to estimate a sequential model of search and, in a later paper (see Kim, Albuquerque and Bronnenberg, 2017), they extend their approach to take advantage of search and purchase data, which helps identifying consumer search costs. In digital markets such as the online markets for hotels, consumers have the possibility to sort and filter search results. Chen and Yao (2017) incorporate consumers' search refinement decisions in a sequential search model, which is estimated using clickstream data from a hotel booking website.

I have mentioned above how De los Santos, Hortaçsu and Wildenbeest (2012) exploit data on browsing behaviour to test sequential search against simultaneous search. The same authors (see De los Santos, Hortaçsu and Wildenbeest, 2017) relax the assumption that consumers know the utility distribution while they search and show how search and purchase data can be used to estimate a model of Bayesian learning.

Honka (2014) also uses consumer search and purchase data to separate the role of search and switching costs in creating inertia in the U.S. auto insurance

[24] Chen and Zhang (2016) identify novel effects of firm entry on consumer search incentives and, in a model of random search, conclude that entry can be excessive from the point of view of consumer welfare. Fishman and Levy (2015) study how search costs affect the incentives to invest in quality in a model of random search with infinitely many firms; they find that the effect is ambiguous. Moraga-González and Sun (2018) focus on the efficiency of market equilibrium and provide conditions under which quality investment can be excessive or insufficient from the point of view of social welfare maximization.

market. She uses the simultaneous search framework of Chade and Smith (2006) and finds that both search and switching costs are significant. Search costs, however, appear to affect inertia much more strongly than switching costs do. This empirical result is in line with Wilson (2012).

Ursu (2017) is another interesting contribution using search and purchase data. She exploits a natural experiment to identify the causal effect of search engine ranking position. She finds that ranking position affects the probability of receiving clicks, but not the probability of selling. More interestingly, she shows that the Expedia ranking is not utility-maximizing, a result somewhat in line with Anderson and Renault (2017).

Classical research domains in industrial organization such as collusion theory and merger analysis have benefited or may benefit from the recent developments in directed search theory. For example, Petrikaitė (2015) investigates the stability of collusion in search costs environments. She concludes that with differentiated products higher search costs make cartels more stable. She studies this problem within the context of a random search model. However, the incentives to deviate clearly depend on whether consumers observe the deviation prices before search or not. Extending her work by allowing for price-directed search would help clarify further the role of search costs in collusion theory.

Moraga-González and Petrikaitė (2013) study mergers in the classical price competition environment with differentiated products. They show that price coordination is not profitable for the firms. The reason is that consumer search is directed and if consumers expect the merged entity to charge higher prices they rather visit it last. By contrast, if the merging firms start stocking the products of the constituent firms after the merger, then the merger becomes incentive-compatible and can even be welfare improving due to better matching between consumers and products.

Price discrimination is another area in which explicitly acknowledging that markets are not frictionless leads to new and fruitful insights. In digital markets in particular, retailers are tracking consumer search behaviour by inserting "cookies" in consumer browsers. Some online retailers offer price discounts when they "observe" that a consumer is going to leave the retailer's website. There have also been allegations that online sellers raise their prices when they "see" a consumer returning to its website. Armstrong and Zhou (2016) study the incentives retailers have to engage in these pricing practices in a duopoly version of Wolinsky's (1986) consumer search model. They show an individual firm always has an incentive to offer buy-now discounts and, under some conditions, to make exploding offers. The use of buy-now discounts or exploding offers reduces social welfare because fewer consumers buy and those who do buy stop searching too early and thus get poorly matched to products.

In the context of a multiproduct firm, Petrikaitė (2018) demonstrates that a monopolist can benefit from the existence of search costs (and therefore has incentives to invest in creating them) to inspect the array of products sold by the firm. She shows that in such a case the monopolist's pricing policy consists of a decreasing sequence of prices. This departure from the symmetric pricing policy increases the profits of the firm at the expense of consumers because of screening. Suppose the seller sells two (substitute) products. Buyers who find the first product good enough do not find it worth to continue searching the next product, which gives the seller market power over these consumers. Consumers who dislike the first product continue searching the next product, which is offered at a lower price. Interestingly, this practice is also profitable under oligopoly.

Finally, I should like to mention the topic of vertical relations and vertical restraints. Only recently has the literature started to incorporate consumer search into vertical relationships. Janssen and Shelegia (2015) argue that it is natural to believe that in markets where consumers need to search across retailers to find acceptable prices, they are likely to ignore the wholesale price as well. In this situation, they show that, relative to the well-known double-marginalization problem, a manufacturer has enhanced incentives to raise its price, thereby worsening even further the market outcome. Another interesting paper is by Wang and Wright (2017), who study a model in which consumers can search for satisfactory products via platforms or directly. Platforms lower search costs but charge commission fees that create a double marginalization problem. This problem manifests itself most crudely due to "showrooming", that is, the possibility that consumers search on the platform but at the moment of buying switch to the direct channel to benefit from lower prices. Wang and Wright study the welfare effects of price parity clauses, that is, contractual clauses imposed by platforms that require the sellers participating in a platform to not quote lower prices elsewhere. These clauses have been the subject of recent policy investigations by the European Commission. Because search in platforms is predominantly directed, the policy conclusions of Wang and Wright would benefit from a further investigation into the role of directed search within platforms.

To end, consumer search theory has already existed for three or four decades but, interestingly enough, it is nowadays more alive than ever before. This is due, on the one hand, to the development of digital markets, not just because the Internet has made researchers aware that search frictions are an important element that cannot be left out of their models, but also because it is now understood that existing models have to be adapted to better capture the features of online markets. On the other hand, this is due to the richness of data that has become available thanks to the Internet. This richness of data,

in particular search and purchase data, has allowed researchers to engage into new empirical challenges. Due to the limited space, in this chapter I have only been able to discuss some of these developments.[25] Digital markets remain highly innovative and there will surely be many more excellent theoretical and empirical consumer-search related contributions in the years to come.

BIBLIOGRAPHY

Anderson, S. P., and R. Renault (1999), "Pricing, product diversity, and search costs: a Bertrand-Chamberlin-Diamond model," *The RAND Journal of Economics:* 719–735.

— (2016), "Search direction: position externalities and position bias," unpublished manuscript: 177-224.

— (2018), "Firm pricing with consumer search," in L. G. Corchón and M. A. Marini (Eds.), *Handbook of Game Theory and Industrial Organization, Vol. II,* Edward Elgar Publishing: 177-224.

Arbatskaya, M. (2007), "Ordered search," *The RAND Journal of Economics,* 38(1): 119–126.

Armstrong, M. (2017) "Ordered consumer search," *Journal of European Economic Association,* 15(5): 989–1024.

Armstrong, M.; Vickers, J., and J. Zhou (2009), "Prominence and consumer search," *The RAND Journal of Economics,* 40(2): 209–233.

Armstrong, M., and J. Zhou (2011) "Paying for prominence," *The Economic Journal,* 121: 368-395.

Athey, S., and E. Glenn (2011), "Position auctions with consumer search," *The Quarterly Journal of Economics,* 126: 1213–1270.

Baye, M. R.; Morgan, J., and P. Scholten (2004), "Temporal price dispersion: evidence form an online consumer electronics market," *Journal of Interactive Marketing,* 18(4): 101-115.

— (2004), "Price Dispersion in the small and in the large: evidence from an Internet price comparison site," *The Journal of Industrial Economics,* 52(4): 463-496.

[25] The avid reader is recommended to continue this thread by reading the excellent surveys of Anderson and Renault (2018) and Armstrong (2017).

Baye, M. R., and J. Morgan (2001), "Information gatekeepers on the Internet and the competitiveness of homogeneous product markets," *American Economic Review,* 91(3): 454-474.

Berry, S. (1994), "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics,* 25(2): 242-262.

Berry, S.; Levinsohn, J., and A. Pakes (1995), "Automobile prices in market equilibrium," *Econometrica,* 63: 841–890.

Chade, H., and L. Smith (2006), "Simultaneous search," *Econometrica,* 74: 1293-1307.

Chen, Y., and Ch. Chuan (2011), "Paid placement: advertising and search on the internet," *The Economic Journal,* 121: F309-F328.

Chen, Y., and T. Zhang (2016), "Entry and welfare in search markets," *The Economic Journal,* 128: 55-80.

Chen, Y., and S. Yao (2017), "Sequential search with refinement: Model and application with click-stream data," *Management Science,* 63(12): 4345-4365.

Choi, M.; Dai, A. Y., and K. Kyungmin (2017), "Consumer search and price competition," *Econometrica,* forthcoming.

De los Santos, B. (2018), "Consumer search on the Internet," *International Journal of Industrial Organization,* 58: 66-105.

De los Santos, B.; Hortaçsu, A., and M. R. Wildenbeest (2012), "Testing models of consumer search using data on web browsing and purchasing behavior," *The American Economic Review,* 102(2): 2955–2980.

— (2017), "Search with learning for differentiated products: evidence from E-commerce," *Journal of Business & Economic Statistics,* 35: 626-641.

De los Santos, B., and S. Koulayev (2013), "Optimizing click-through in online rankings for partially anonymous consumers," *Marketing Science,* forthcoming.

Diamond, P. A. (1971), "A Model of Price Adjustment," *Journal of Economic Theory,* 3: 156–168.

Ding, Y., and T. Zhang (2018), "Price-directed consumer search," *International Journal of Industrial Organization,* 58: 106-135.

Edelman, B.; Ostrovsky, M., and M. Schwarz (2007), "Internet advertising and the generalized second price auction: selling billions of dollars worth of keywords," *American Economic Review,* 97(1): 242–59.

Ellison, G., and A. Wolitzky (2012), "A search cost model of obfuscation," *The RAND Journal of Economics,* 43(3): 417–441.

Ershov, D. (2018), "The effects of consumer search costs on entry and quality in the mobile app market," unpublished manuscript.

Fishman, A., and N. Levy (2015), "Search costs and investment in quality," *The Journal of Industrial Economics,* 63: 625-641.

Fishman, A., and D. Lubensky (2018), "Search prominence and return costs," *International Journal of Industrial Organization,* 58: 136-161.

Giulietti, M.; Waterson, M., and M. R. Wildenbeest (2014), "Estimation of search frictions in the British electricity market," *Journal of Industrial Economics,* 62: 555-590.

Haan, M.; Moraga-González, J. L., and P. Vaiva (2017), "A Model of directed consumer search," unpublished manuscript.

Hong, H., and M. Shum (2006), "Using price distributions to estimate search costs," *The RAND Journal of Economics,* 37: 257-275.

Honka, E. (2014), "Quantifying search and switching costs in the U.S. auto insurance industry," *The RAND Journal of Economics,* 45: 847–884

Honka, E., and P. Chintagunta (2016), "Simultaneous or sequential? search strategies in the US auto insurance industry," *Marketing Science,* 36(1): 21-42.

Hortaçsu, A., and Ch. Syverson (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: 403–456.

Janssen, M. C. W.; Moraga-González J. L., and M. R. Wildenbeest (2005), "Truly costly sequential search and oligopolistic pricing," *International Journal of Industrial Organization,* 23(5): 451–466.

Janssen, M. C. W., and A. Parakhonyak (2014), "Consumer search markets with costly revisits," *Economic Theory,* 55: 481–514.

Janssen, M. C. W., and S. Shelegia, "Consumer search and double marginalisation," *The American Economic Review,* 105(6): 1683-1710.

Kim, J. B.; Albuquerque P., and B. J. Bronnenberg (2010), "Online demand under limited consumer search," *Marketing Science,* 29: 1001–1023.

— (2017), "Online demand under limited consumer search," *Management Science,* 63(11): 3911–3929.

Kohn, M. G., and S. Shavell (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.

McCall, J. J. (1970), "Economics of information and job search," *Quarterly Journal of Economics,* 84: 113–126.

Moraga-González, J. L., and V. Petrikaitė (2013), "Search costs, demand-side economies and the incentives to merge under Bertrand competition," *The RAND Journal of Economics,* 44: 391-424.

Moraga-González, J. L.; Sándor, Z., and M. R. Wildenbeest (2013), "Semi-nonparametric estimation of consumer search costs," *Journal of Applied Econometrics,* 28: 1205-1223.

— (2017a), "Nonsequential search equilibrium with search cost heterogeneity," *International Journal of Industrial Organization,* 50: 392-414.

— (2017b), "Prices and heterogeneous search costs," *The RAND Journal of Economics,* 48(1): 125-146.

— (2017c), "Consumer search and prices in the automobile market," unpublished manuscript.

Moraga-González, J. L., and Y. Sun (2018), "Quality provision and consumer search," unpublished manuscript.

Moraga-González, J. L., and M. R. Wildenbeest (2008), "Maximum likelihood estimation of search costs," *European Economic Review,* 52: 820-848.

Morgan, P., and R. Manning (1985), "Optimal search," *Econometrica,* 53: 923–944.

Mortensen, D. (1970), "Job search, the duration of unemployment, and the Phillips curve," *The American Economic Review,* 60(5): 847-62.

Nevo, A. (2001), "Measuring market power in the ready-to-eat cereal industry," *Econometrica,* 69: 307–342.

Petrikaitė, V. (2016), "Collusion with costly consumer search," *International Journal of Industrial Organization,* 44: 1-10.

— (2018), "Consumer obfuscation by a multiproduct firm," *The RAND Journal of Economics,* Vol. 49: 206-223.

Pratt, J: W.; Wise, David A., and R. Zeckhauser (1979), "Price differences in almost competitive markets," *Quarterly Journal of Economics,* 93(2): 189-211.

Rhodes, A. (2011), "Can prominence matter even in an almost frictionless market?," *The Economic Journal,* 121: F297–F308.

Rob, R. (1985), "Equilibrium price distributions," *Review of Economic Studies,* 52: 452–504.

Rothschild, M. (1973), "Models of market organization with imperfect information: a survey," *Journal of Political Economy,* 81(6): 1283-1308.

Shelegia, S., and D. Garcia, "Consumer search with observational learning," *The RAND Journal of Economics,* forthcoming.

Sanches, F.; Silva Junior, D., and S. Srisuma, "Minimum distance estimation of search costs using price distribution," *Journal of Economics and Business Statistics,* forthcoming.

Stahl, D. O. (1989), "Oligopolistic pricing with sequential consumer search," *The American Economic Review,* 79(4): 700–712.

— (1989), "Oligopolistic pricing with heterogeneous consumer search," *International Journal of Industrial Organization,* 14: 243-268.

Stigler, G. J. (1961), "The Economics of information," *The Journal of Political Economy,* 69: 213-225.

Stiglitz, J. E. (1989), "Imperfect information in the product market," in Schmalensee, R., and Willig, R. D. (Eds.), *Handbook of Industrial Organization,* Vol. 1: 769-847.

Ursu, R. (2017), "The power of rankings: quantifying the effect of rankings on online consumer search and purchase decisions," *Marketing Science,* forthcoming.

Varian, H. R. (1980), "A model of sales," *The American Economic Review,* 70(4): 651–659.

— (2007), "Position auctions," *International Journal of Industrial Organization,* 25(6): 1163–1178.

Weitzman, M. L. (1979), "Optimal search for the best alternative," *Econometrica,* 47(3): 641-654.

Wilson, C. M. (2010), "Ordered search and equilibrium obfuscation," *International Journal of Industrial Organization,* 28(5): 496–506.

— (2012), "Market frictions: A unified model of search costs and switching costs," *European Economic Review,* 56: 1070–1086.

Wolinsky, A. (1986), "True monopolistic competition as a result of imperfect information," *The Quarterly Journal of Economics,* 101(3) 493–511.

Wang, Ch., and J. Wright (2017), "Search platforms: showrooming and price parity clauses," unpublished manuscript.

Wright, R.; Kircher, P.; Julien, B., and V. Guerrieri (2017), "Directed search: a guided tour," unpublished manuscript.

Zhou, J. (2011), "Ordered search in differentiated markets," *International Journal of Industrial Organization,* 29(2): 253–262.

— (2014), "Multiproduct search and the joint search effect," *The American Economic Review,* 104(9): 2918–2939.

PART III

New Digital Business Models

CROWDFUNDING: WHAT DO WE KNOW?[1]

Carlos BELLÓN
Pablo RUIZ-VERDÚ

Abstract

Crowdfunding is a new form of financing that takes place through online platforms and involves the participation of a large number of contributors. Because of crowdfunding's potential to aggregate the information dispersed among many potential consumers or investors, crowdfunding is regarded by many as a revolutionary way of financing new ventures. At the same time, crowdfunding is touted as a way to democratize investment in entrepreneurial firms, which regulation has kept mostly outside the reach of small investors. In this article we survey the role that crowdfunding may play as a source of financing for entrepreneurial firms. To do so, we review the existing theoretical and empirical work on crowdfunding and discuss the ways in which crowdfunding differs from alternative sources of financing, such as angel investing or venture capital, and whether it may replace or complement these other financing sources. We also describe how crowdfunding has been regulated so far and identify key open questions in the regulatory debate.

Key words: Crowfunding, equity crowdfunding, reward crowdfunding, entrepreneurial finance, regulation.

JEL classification: G32, G38.

[1] The authors would like to gratefully acknowledge the financial support provided by the Comunidad de Madrid through grant H2015/HUM-3417. Pablo Ruiz-Verdú also acknowledges the support of Spain's Ministry of Economy and Competitiveness (through grant ECO2015-69615-R) and of FEDER (UNC315-EE-3636).

I. INTRODUCTION

Crowdfunding (CF) has been described as a revolutionary new method to finance entrepreneurial ventures and been touted as a way to democratize investment in early-stage firms, a kind of investment previously reserved to institutional investors and wealthy individuals. As its name indicates, CF involves the contributions of a large number of investors, and, as a result, the development of CF has been accompanied by regulatory changes aimed at promoting CF's benefits while minimizing the potential risks for small investors of investing in entrepreneurial firms. In this paper, we describe the phenomenon of CF as a form of entrepreneurial financing, review the academic research related to it, and discuss the regulatory concerns that CF raises.

In this review, we frame CF within the broader context of entrepreneurial finance, in order to understand how it addresses the three basic problems faced by any form of early-stage financing: the large degree of uncertainty about the project being funded, the fact that the information about the prospects of the project is asymmetrically distributed among the entrepreneur and potential investors, and the possibility that the entrepreneur does not use the funds in the interest of investors. We argue that CF deals with these problems in ways very different from those of traditional sources of entrepreneurial finance, such as venture capital (VC), angel investing, or bank financing, and review existing research to better understand these differences. In so doing, we seek to shed light on the question of whether CF can serve as a source of funding for projects which would otherwise have no or very costly access to financing, and identify the features of both projects and CF design that make CF attractive as a financing source.

We would like our review to be of interest to non-specialists who want to understand what CF is all about. At the same time, we hope to provide an integrated view of the academic literature to researchers interested in CF and help them identify promising research questions. Finally, we hope to inform the regulatory debate, both by providing regulators with an overview of the academic research and by pointing researchers toward research questions that matter to regulators.

It is important to note that we do not consider in this review all the financing forms that are sometimes described as CF. Although there is no generally accepted definition of CF, it is fair to say that the term is used to describe a fund raising process if it solicits contributions from a large number of individuals through the internet.[2] However, these two features,

[2] Although there are many examples of fund-raising campaigns enlisting large numbers of donors (the Joseph Pulitzer-led campaign in 1885 to fund the construction of a plinth for the Statue of Liberty with contributions from 160,000 donors being an oft-cited example, BBC News, 2013), the term CF is normally used when the fund raising process takes place online.

namely the participation of a large number of contributors (or *backers*) and the use of the internet, characterize a broader array of financing practices. In this paper, we will focus on the financing of early-stage business ventures through the two practices to which the term CF is most commonly applied and have received most attention in the academic literature: reward and equity CF. In reward CF, contributors provide funding to an entrepreneur in exchange for a non-monetary reward, which is often the good or service for whose production the entrepreneur is raising money, but which can range from a simple thank-you message to special versions of the good with customized add-ons. In equity CF (also known as securities-based CF or crowdinvesting), contributors provide funds to the entrepreneur in exchange for a monetary return. In both reward and equity CF, the process takes place through an online CF platform. This narrow focus on reward and equity CF leaves out funding practices that do not finance business ventures, such as peer-to-peer (P2P) consumer lending, donation CF, or real estate CF. We also leave out of our discussion P2P lending to businesses to keep this review manageable.[3] Although the distinction between P2P business lending and some forms of equity CF is not clear-cut in theory, the evolution of the P2P business lending model has led P2P lending platforms to play roles distinct from those of equity CF platforms and more akin to those of traditional lending intermediaries (see Bachman *et al.,* 2011, for a review of the P2P literature). Finally, we will not cover in our discussion very recent phenomena, such as Initial Coin Offerings (ICOs) or Tokenized Asset Offerings (TAOs), which are forms of crowfdunding crypto-currency projects, and which have experienced explosive growth in the last few years (see Burniske and Tatar, 2018, for a detailed introduction).

We should also mention that CF has attracted the attention of researchers from many different fields. Although we will not limit our discussion to contributions from economics and finance, we will focus in our review of the theoretical literature on the contributions from these fields.

There are a few other introductions to CF and surveys of the CF literature. Agrawal, Catalini and Goldfarb (2014) and Belleflamme, Omrani and Peitz (2015) only cover the very early literature on CF, but are still nice introductions to the phenomenon and to some of the economic questions that it raises. Moritz and Block (2016) and Wallmeroth, Wirtz and Groh (2018) provide more recent reviews of the literature. In both cases, the authors aim to provide a comprehensive account of everything that has been published on CF, so they complement this survey. Short *et al.* (2017) and McKenny *et al.* (2017) review

[3] Crowdfunding can be considered to be a part of a broader phenomenon that the Cambridge Center for Alternative Finance terms "alternative finance," and which comprises all "technology-enabled online platforms (or channels) that act as intermediaries in the de mand and supply of funding to individuals and businesses outside the traditional banking system" (Ziegler *et al.,* 2017: 20).

the work on CF in management and entrepreneurship journals and identify areas for future research.

The paper is organized as follows. In Section II, we describe the problems of asymmetric information and opportunism that are prevalent in the financing of entrepreneurial firms, how both traditional financing forms and CF address these problems, and provide evidence to evaluate the current relevance of CF as a funding source, its evolution, and the types of projects that it finances. We review the theoretical literature in Section III and the empirical literature in Section IV. In Section V, we discuss the regulatory concerns that CF raises and how CF has been regulated so far. Finally, we provide some concluding remarks in Section VI.

II. FINANCING NEW VENTURES

1. The Basic Problems

There are two basic problems that need to be addressed when providing financing to a firm to carry out an investment project.

Uncertain project quality. The first problem is that the quality of the project is uncertain. This uncertainty could be due to uncertainty about demand, the ability of the firm to complete the project as proposed, or the trustworthiness of the firm's managers. Most often, different parties have different information about the determinants of project quality. Would-be buyers are informed about their own demand for the product and the firm's managers are better informed about their trustworthiness and, often, about the technical viability of the project and its cost.

Conflicts of interest and moral hazard. The second problem is that the interests of the investors and those of the firm's managers may not be aligned. For example, the firm's managers may prefer to invest in certain projects that they find intrinsically rewarding or that may give them visibility. Managers may also have a preference for control and, thus, resist being replaced by others who are better able to pursue the investors' goals. Managers may also pay themselves excessive compensation, shirk on their duties, devote too much of their time to activities that they find appealing but which may not be optimal for investors, or even divert the company's funds for their personal use.

2. The Basic Elements of a Financing Form

Financing forms differ in the way they address these two main problems by means of initial screening, monitoring and advising, and the allocation of decision rights and returns.

Initial screening. Prior to providing the funds, the potential investors or some intermediaries may investigate the project and the managers to obtain information about the value of the former or the ability or trustworthiness of the latter. Investors may also structure the financing deal so as to induce managers to reveal information through their very acceptance of the deal, or one of the deals, proposed by investors.

Monitoring and advising. Once financing takes place, investors or intermediaries may monitor the manager in order to avoid opportunistic behavior or provide advice to the manager.

Decision rights. To allow investors to control the manager's behavior, the financing form may limit the manager's choices or provide certain decision rights to investors. One possible way to allocate decision rights is by means of staging; that is, by providing only partial financing, so that the initial or future investors effectively have the right to discontinue the project.

Returns. As an alternative to intervention, the allocation of returns to the manager can also be designed so as to provide the manager the incentives to act in the investors' interests.

Formal contracts and informal agreements. The firm and its investors may write a formal contract that specifies these four elements explicitly or may tacitly agree to some of them.

Investment projects may differ in how uncertain they are and in whether the uncertainty stems mostly from the demand for the product, the manager's ability or trustworthiness, or the technical feasibility or cost of the project. Projects may also differ in their size, in the non-pecuniary benefits and costs that they may generate for firms and investors, and in the severity of the conflict of interest between investors and managers. Different financial forms will be differently adapted to the characteristics of the investment projects, so one expects that some financial forms will be used for some kinds of projects but not for others.

3. Sources of Financing for New Ventures

New ventures are likely to exhibit greater uncertainty and opacity than more established firms, since early-stage firms do not have a record that can be analyzed to determine the value of their project or the ability of their managers. Moreover, in many cases, new ventures bring new goods to the market, whose prospects may be difficult to evaluate.

The financing forms used to finance early-stage firms should thus be especially adapted to their uncertainty. We briefly describe the main forms used in the financing of early-stage firms and then describe in greater detail the two main forms of CF.

Before doing so, we note that we will use the term *early-stage financing (or entrepreneurial or new venture financing*) to encompass the financing of several stages in the early life of a company. Although the terminology is not used uniformly, these stages can be described as: *seed* financing, which finances the very first activities in the development of the project, like assembling the management team or early product development; *startup* or *early stage* financing, which finances product development, marketing, and initial operations, mostly before the firm generates revenues; *growth*; *expansion* financing; and *mezzanine* financing, which refer to further stages in the development of the business.

3.1. Venture Capital

The term venture capital (VC) financing refers to the financing to early-stage firms provided by professional investors who work on behalf of institutions or wealthy individuals. VC funds generally syndicate their investments, with several funds contributing to a round of financing (see Rin, Hellmann and Puri, 2013; Kaplan and Strömberg, 2003; or Kaplan and Strömberg, 2004 for general references about VC).

VC funds invest mostly in businesses with a high growth potential. Although they provide some seed financing, they invest primarily in companies in the startup and growth stages. VC financing deals are generally above US$1 million, with typical deal sizes in the range from US$1 to 2 million for seed stage deals, up to US$20 to 60 million for later stage deals, and even more than US$100 million in the case of so-called mega-rounds (PwC/CB Insights, 2018).

VC firms employ a variety of methods to select the companies they finance, which range from going through business plans submitted by entrepreneurs to referrals by founders or employees of former portfolio companies or lawyers. In any case, before a VC fund decides to invest in a company, it performs extensive due diligence, which, among other things, involves analyzing business plans thoroughly, meeting several times with founders, or consulting references.

VC firms perform several monitoring and advising roles in their portfolio companies. Thus, VC firms may help recruit senior management and board

members, sit at the board, and may help companies obtain additional financing. VC partners frequently interact with the founders and employees of portfolio companies.

VC financing is characteristically provided in stages, either by means of *ex ante* staging (the funding within a financing round is staged condition ally on achieving performance goals) or *ex post* (between rounds) staging, by ensuring that initial financing is not sufficient to cover the firm's financing needs until exit (Kaplan and Strömberg, 2003). Staging reduces the uncertainty faced by investors and provides incentives to entrepreneurs, who have to perform well to secure funding in subsequent rounds.

VC funds use different securities to finance early stage firms. The most common one is some form of convertible preferred equity, which combines a debt-like preferred security with an option to convert into an equity-like security. If the exit value (*i.e.,* the value of the firm at the time it goes public or is acquired by another firm) is low, VCs obtain the preferred terms, but convert to common equity if the exit value is high. Very importantly, financing contracts usually contain several clauses that restrict the actions that the entrepreneur can take (negative covenants), confer decision rights to VCs, and protect the VCs stake in the firm from being diluted in subsequent financing rounds.

3.2. Angel Investors

Angel investors are wealthy individuals, often with experience as entrepreneurs or managers, who invest their own money in private companies not managed or owned by family or friends. Angel investors often invest individually, but they also invest in formal or informal groups or syndicates or join angel investor networks.

Angel investors get to know about potential deals through informal personal networks and, recently, also through more formal networks. Al though there is heterogeneity in the angel investors' screening process, which has become more formalized in the case of investments by angel groups, it generally involves several meetings with the entrepreneur and the performance of due diligence (OECD, 2011; Shane, 2005).

There is also heterogeneity in the use of financing contracts, although for traditional individual angel investors, common stock with few contractual protections is prevalent (Wong, Bhatia and Freeman 2009; Goldfarb *et al.,* 2013; Shane, 2005; DeGennaro, 2012).

In terms of their monitoring and advising role, the degree of involvement of angel investors ranges from mostly passive investment to becoming CEOs (Shane, 2005; DeGennaro, 2012; Wong, Bhatia and Freeman, 2009; Goldfarb *et al.,* 2013; OECD, 2011). Even if angel investors often do not sit at the board, they interact informally with managers, help form the management team, provide operational assistance, and help to procure further financing. Very often, angel investors invest only in firms that are geographically close, so as to be able to provide informal consultation and monitoring.

Although some angel investors make use of contracts with protections similar to those used by VCs, to a large extent, angel investors substitute personalized monitoring, influence, and implicit incentives for the more formal protections and control rights provided by VC contracts as vehicles to incentivize and monitor entrepreneurs.

Angel investors typically invest in earlier stages than VCs, particularly in seed financing. Individually and as groups, investment size tends to be smaller than in the case of VC, in the range of US$100,000 to 2 million, a range that is generally not covered by VC (Ibrahim, 2008; Wong, Bhatia and Freeman, 2009; Shane, 2005, 2012; Goldfarb *et al.,* 2013).

3.3. IPOs

The Initial Public Offering (IPO) is the issuance of shares in the firm to the general public through listing in a regulated stock exchange. More than an alternative source of new venture finance, the IPO is considered as one of the two exit strategies (the other being the acquisition by another firm) for investors in early-stage companies to liquidate their stakes. Thus, in their study of US IPOs between the years 1993 and 2003, Bradley *et al*. (2009) document that the median company age at the IPO was 7 years and the median size was US$42.5 million.

An IPO involves a two-stage screening process. Investment banks, acting as gatekeepers, first decide whether to underwrite the equity issue. Since investment banks put their own capital and reputation at risk, they will accept to serve as underwriters if they have a strong faith in the firm's valuation. If they are uncertain about the level or volatility of the price of the newly issued stock they may elect to advise the firm on a best-efforts only basis (with a substantially reduced fee). Finally, if they do not believe the firm is capable to comply with the increased transparency and professionalization requirements of a public company they will not take part in the listing.

Once taken on as a client, investment banks help the company prepare to comply with all the reporting that is necessary for a public offering of shares. The regulation of IPOs requires the firm to disclose (through the *Prospectus*) a substantial amount of information about its governance, management, business model, and accounts, which should be audited. It also requires the firm to register the securities with the relevant authorities (such as the SEC in the US) and to commit to communicate all relevant information promptly to the market under penalty of legal sanction. The explicit costs of this process can reach hundreds of thousands of dollars (Bradford, 2012), excluding underwriting fees, which can themselves be in the millions (GAO, 2000, estimate underwriting fees to be around ten percent of total proceeds for an IPO). Furthermore, from the moment of listing, public companies experience strong monitoring through the work of investment analysts, institutional investors, regulators and the secondary market itself via prices.

3.4. Bank Lending

Firms looking for bank lending are typically screened by a loan officer, who examines the applicant's business model, financial projections, and assets. Since bank financing generally takes the form of debt, banks are often unwilling to lend to firms with significant downside risk (*i.e.,* with a significant risk of default) or require those firms to post collateral. Often, this means that loans are undertaken in the name of the entrepreneur herself rather than the business (Cole and Sokolyk, 2017).

Banks may monitor borrowers by means of periodic inspections of the firm's accounts (and possibly offices) by the loan officer. Although banks provide some advice to their borrowers, the scope of such advice is much more limited than that provided by VCs or angel investors. Despite not owning equity in the firm, banks are able to exercise decision rights by a careful selection of loan maturities and through explicit covenants in the loan agreements.

3.5. Reward Crowdfunding

In reward CF, the entrepreneur posts a description of her project on an online CF platform or portal (such as US portals *Kickstarter* or *Indiegogo*, UK portal *Crowdfunder*, or French portals *Ulule* or *KissKissBankBank*) upon approval by the portal, which is expected to perform minimal or no due diligence. Table 1 lists the most popular platforms (both for reward and equity CF) worldwide as measured by internet traffic and Table 2 displays some key characteristics of the top platforms.

TABLE 1

TOP CROWDFUNDING SITES
(Top CF platforms based on number of visits as reported by similarweb.com)

Site Name	*Type*	*Country*
Kickstarter	Reward	US
Indiegogo	Reward	US
Angel.co	Equity[a]	US
Ululue	Reward	France
Pledgemusic	Reward	UK & US
SeedandSpark	Reward	US
CircleUp	Equity[a]	US
Kickante	Reward	Brazil
Crowdcube	Equity	UK
Seedrs	Equity	UK
Seedinvest	Equity	US

Notes: [a]Accredited investors only.
Source: https://crowdfundingpr.wordpress.com/2016/05/01/top-100-crowdfunding-sites-in-the-united-states-europe-asia- south-america-africa-and-other-global-markets-in-2016/. Accessed April, 2018.

TABLE 2

KEY STATISTICS ON SELECTED CROWDFUNDING PLATFORMS

Site	*Total Raised US$ mn*	*Platform Fee (%)*	*Payment Fee*	*Comments*
Kickstarter	3,000	3-5	3-5%	
Indiegogo	1,000	5	3-5%+US$ 0.30	
Seedrs	450	6	0.5%	GBP2,500 completion fee
Crowdcube	400	7	0.5%-2.9%	Payment fee depends on location
Wefunder	50	up to 7		Investors are charged 2% with a maximum of US$75. For Reg D, Wefunder charges up to 20% carried interest.

Notes: GBP converted to US$ using an average exchange rate of 0.770.
Sources: (Accessed April, 2018): (a) www.kickstarter.com/help/faq, (b) support.indiegogo.com, (c) www.seedrs.com/learn (d) help.crowdcube.com, (e) www.wefunder.com/faq, (f) Reg. D offerings are for ac- credited investors only.

The description of a reward CF campaign includes information on the good or service that the entrepreneur (also referred to as the *sponsor*) plans to develop with the funds raised, a funding goal, and a menu of pledge levels (that is, possible contribution levels) with associated rewards. The menu of rewards typically includes the delivery of the product, often in different versions or with customizations or add-ons. However, there is a wide array of rewards, which

often include some form of public appreciation by the entrepreneur (such as including the backer's name in a list of contributors) or promotional materials such as t-shirts or the invitation to events.

Reward CF campaigns accept contributions for a limited period of time (often one to two months). During the campaign, potential backers can observe the amount of money contributed so far and may observe additional information about the distribution of pledges and even the identity of the pledgers. Reward CF portals generally offer public channels of communication between backers and sponsors.

Reward CF campaigns generally come in one of two formats, All or Nothing (AoN) campaigns, in which the contributions made by backers are refunded to them if a funding threshold is not met, and Keep It All (KIA) campaigns, in which the entrepreneur keeps all the money contributed by backers irrespectively of the total amount contributed. Some platforms (*e.g.,* Indiegogo) allow entrepreneurs to choose the campaign format, whereas others (*e.g.,* Kickstarter) offer only one format.

If the project is funded, the entrepreneur is expected to deliver the rewards by the stated deadlines and to keep an open communication with backers, but has no other obligations towards backers or the portal.

In reward CF, the screening is carried out almost solely by potential backers, who use the information provided by the entrepreneur in the campaign's site, the comments by other backers in the platform's boards, and, commonly, additional information gathered from online social media to evaluate the entrepreneur and the project. Importantly, since the campaign takes place over time and platforms provide information about the history of the contributions, potential backers can also use this information to determine their funding decisions.

In stark contrast with VC and angel investor funding, many backers will have no contact with the entrepreneur (although "family and friends" are frequent backers), the amount of due diligence performed by potential backers is generally small, the contract between backers and sponsors provides the former no control rights and essentially no explicit contractual protections beyond those provided by the general regulatory framework (see Section V for a discussion of this regulatory framework), and backers do not have access to the entrepreneur except through the platform's message boards or online social media. The mechanisms by which reward CF addresses the problems of screening and motivating entrepreneurs are thus radically different from the ones that characterize traditional forms of entrepreneurial financing. At the

same time, if the product is offered as a reward, reward CF closely resembles the practice of pre-selling, which has a long tradition in consumer markets. However, reward CF differs from pre-selling in that in CF there is generally much greater uncertainty about whether, how, and when the product will be developed. When more symbolic rewards are offered to backers, reward CF approaches donation CF, which, except for the role of the online platform as intermediary, is not essentially different from traditional forms of fundraising for charitable or artistic projects.

3.6. Equity Crowdfunding

Equity CF works in the same way as reward CF, except for two essential differences. The first, defining, difference is that backers in equity CF receive as compensation for their contribution a contractual claim to a monetary return. This contractual claim may take the form of common stock in the company being financed (thus the name equity CF), but often adopts other forms, such as different versions of convertible debt or promises to receive equity in the future (see Section IV.1 for a description of the most common contractual forms). Thus, it is probably more accurate to label this form of CF, securities-based CF or crowdinvesting, but in keeping with usual practice, we refer to it as equity CF. Backers' contracts often include some protections similar to those typical of VC contracts, but protections are generally fewer and weaker than those afforded to VCs.

The other difference with respect to reward CF is that equity CF platforms typically have to perform much more due diligence than reward CF platforms, including, among other things, performing background checks of entrepreneurs or checking that the financial information and other material statements provided by entrepreneurs are correct.

As in the case of reward CF, and in contrast with VC and angel investing, the screening process involves no face-to-face interaction between backers and sponsors, who are not expected to be professional investors, may have no industry experience, and are likely to perform relatively little due diligence. Similarly, equity CF backers typically have little or no formal control rights and no informal levers of influence or channels of advice except those provided by the online communication between backers and sponsors through the CF platform or online social media. Despite these differences, some forms of equity CF are close to the more impersonal and online-mediated form of angel investing characteristic of group angel investing through online angel networks. This kind of angel investing could be even considered a form of equity CF, except that the "crowd" is generally restricted by regulation to be composed of wealthy or sophisticated investors. Equity CF is also formally very similar to IPOs, but it

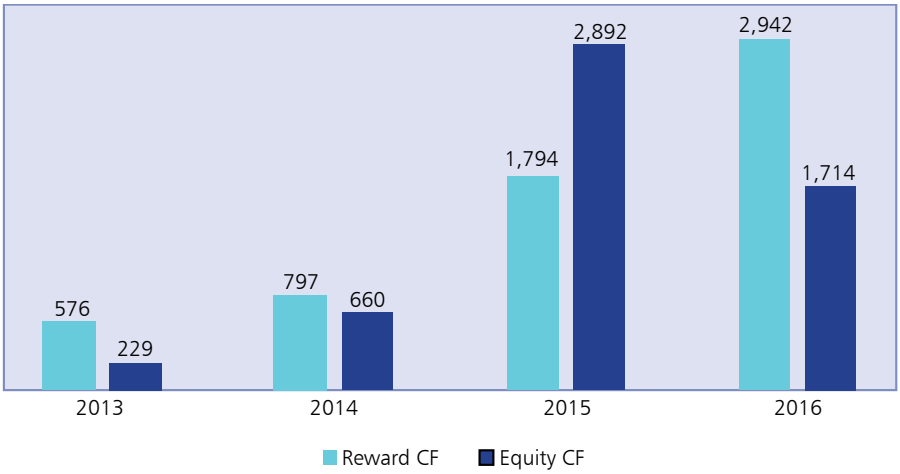
differs from IPOs in that the information disclosure requirements for equity CF are weaker, the amount raised is typically limited to lie below some threshold, there is a very limited secondary market, or none at all, for the securities –which are not traded in regulated exchanges–, the types of securities issued are more varied, and the role of the CF platform is more limited than that of IPO underwriters.

4. How Relevant is Crowdfunding in the Financing of New Ventures?

How prevalent is CF as a source of financing for new business ventures? Before we describe the available evidence about the size of CF in relation

FIGURE 1

EVOLUTION OF CROWDFUNDING MARKET SIZE

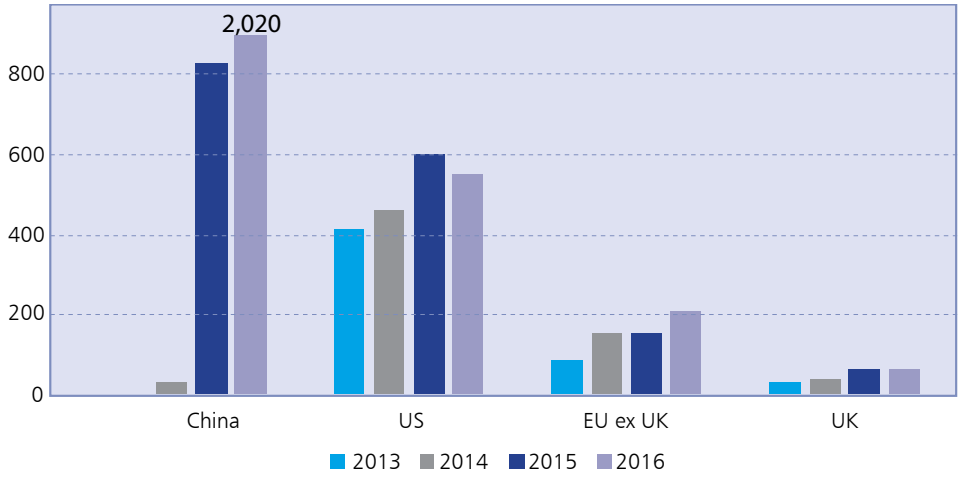


Note: Total worldwide volume raised (in US$ million). Reward CF is defined as including CF projects that provide backers with non-monetary rewards, some of which may be significant enough so as the expense not to be considered a pure donation. Equity CF is defined as including CF projects that provide backers with a monetary return tied to the project's performance; includes platforms targeting accredited investors only, and those that cover the wider public. Angel investing volumes are estimates that include both the visible and non-visible market.

Source: Worldwide volumes computed by adding regional numbers reported for Asia Pacic (Garvey *et al.,* 2017), Africa and the Middle East (Rau *et al.,* 2017), the Americas (Ziegler *et al.,* 2017), Europe (Ziegler *et al.,* 2018) and the UK (Zhang *et al.,* 2017). GBP data translated to US$ at average exchange rate of 0.770 (2016) and 0.681 (2015). Euro data translated to US$ at average exchange rate of 0.940 (2016) and 0.937 (2015).

FIGURE 2

REWARD CROWDFUNDING MARKET SIZE BY REGION



Note: Reward CF volume (in US$ million) for selected regions. The 2016 figure for China is not to scale. Reward CF is defined as including CF projects that provide backers with non-monetary rewards, some of which may be significant enough so as the expense not to be considered a pure donation.

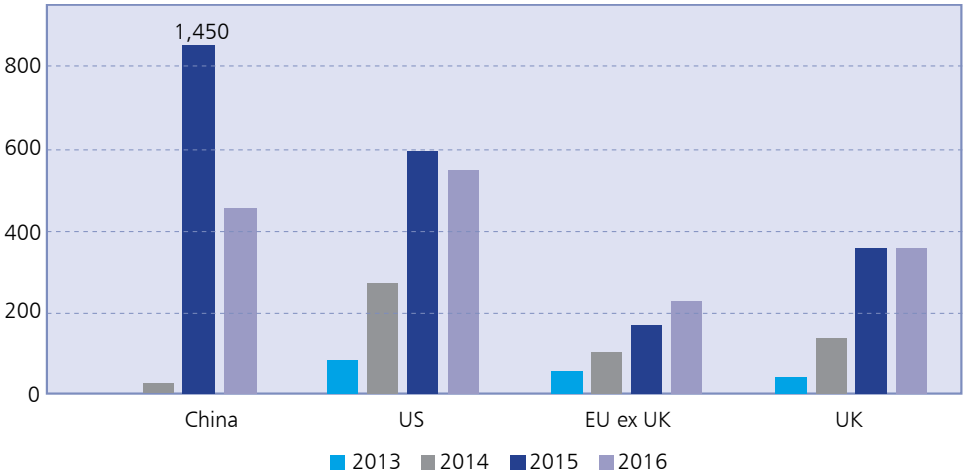
Sources: China (Garvey *et al.,* 2017), US (Ziegler *et al.,* 2017), Europe (Ziegler *et al.,* 2018) and the UK (Zhang *et al.,* 2017). GBP converted to US$ using an average exchange rate of 0.770 (2016) and 0.681 (2015). Euro converted to US$ using an average exchange rate of 0.940 (2016) and 0.937 (2015).

to other forms of entrepreneurial finance, it is important to emphasize that there is no single authoritative source of data on CF and that, in general, the breadth and quality of information on entrepreneurial financing are limited. Therefore, the numbers that we provide should be interpreted with care as broad approximations.

Figure 1 illustrates the development of reward and equity CF worldwide. Even though the beginnings of CF can be dated to the early 2000s, by 2013 reward CF raised more than US$500 million and by 2016 almost US$3 billion worldwide. As Figure 2 shows, reward CF has grown significantly in Europe and the US since 2013 (although at a decreasing rate and with a drop in the US in 2016). However, the growth of reward CF has been explosive in China, which accounts for about two thirds of the total amount raised globally in year 2016. Figure 3 shows that equity CF has grown significantly as well, although with a large drop in 2016 due mainly to changes in the Chinese regulatory environment (Garvey *et al.,* 2017). It is worth noting that the numbers for equity CF include "crowdfunding" restricted to accredited or qualified investors (which is effectively the only type of CF that is allowed in China, and in the US up to

FIGURE 3

EQUITY CROWDFUNDING MARKET SIZE BY REGION

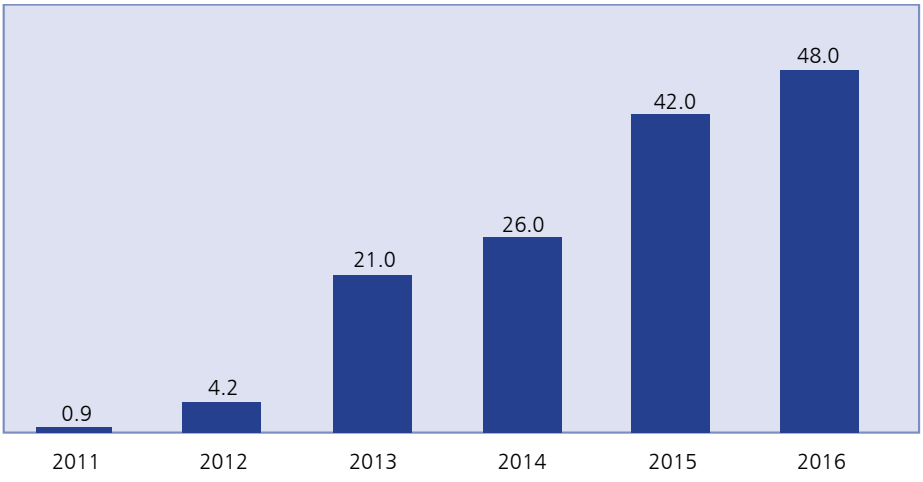


Note: Equity CF volume (in US$ million) for selected regions. The 2015 figure for China is not to scale. Equity CF is defined as including CF projects that provide backers with monetary return tied to the project's performance; includes platforms targeting accredited investors only, and those that cover the wider public.

Sources: China (Garvey *et al.,* 2017), US (Ziegler *et al.,* 2017), Europe (Ziegler *et al.,* 2018) and the UK (Zhang *et al.,* 2017). GBP converted to US$ using an average exchange rate of 0.770 (2016) and 0.681 (2015). Euro converted to US$ using an average exchange rate of 0.940 (2016) and 0.937 (2015).

FIGURE 4

EVOLUTION OF REWARD BASED CROWDFUNDING IN THE UK

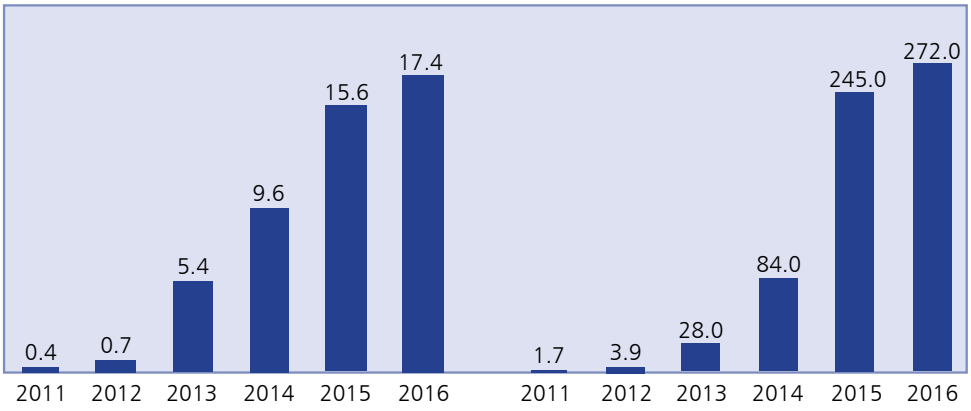


Note: Total volume in GBP million.

Source: Zhang *et al.* (2017).

FIGURE 5

EVOLUTION OF EQUITY CROWDFUNDING IN THE UK



(a) Percent of total seed & venture equity financing (b) Total volume (in GBP million

Source: Zhang *et al.* (2017). VC data sourced from Beauhurst. *Seed* is defined as investment in companies being set up or seeking finance to establish or develop their products further. *Venture* investment is investment in companies with some years of history and in the process of gaining significant traction.

2016). In figures 4 and 5 we also report the evolution of CF in the UK, since it is the economy for which there is a longer time series available.

Table 3 compares the funding volume and number of deals of equity and reward CF with other sources of new venture finance, such as angel investors or VC, and thus allows us to gauge the relative importance of CF as a source of entrepreneurial financing. The UK case is especially noteworthy, since equity CF in this country represented 18.9% of all seed and venture-stage equity investment in 2016,[4] and the number of equity CF deals is 36% of the number of VC deals at the seed or venture stages. British Business Bank (2017) report that CF platforms were involved in 34% of all seed-stage deals (including not only VC, but also angel investor financing) and that they undertook more seed-stage deals (192 deals) than VC (132) in 2016. Although there were about 4,000 reward CF deals funded in the UK in year 2016, the amount raised by reward CF represents less than 20% of the corresponding amount for equity CF.

In the rest of the EU, the total equity CF volume in 2016 represented about 5% of total seed and venture VC funding and was similar to the amount raised by reward CF. For equity CF, Germany (EUR47 million), Sweden (EUR467 million),

[4] A number that is in line with the 17.37% reported by Zhang *et al.* (2017) and that we include in Figure 5.

TABLE 3

MARKET SIZE OF DIFFERENT FINANCING SOURCES FOR NEW VENTURES
(Total volume raised [in us$ miilion] and number of deals by source of new venture finance for selected regions)

	US		*EU ex UK*		*UK*	
	(US$ mn)	*Deals*	*(US$ mn)*	*Deals*	*(US$ mn)*	*Deals*
Reward-based crowdfunding						
2016	551.4[a]	22,056[a]	203.2[b]	12,675[b]	62.3[c]	4,068[c]
2015	601.2[a]	15,820[a]	148.3[b]	32,583[b]	61.7[c]	N.A.[c]
Equity-based crowdfunding						
2016	549.1[a]	637[a]	233.0[b]	724[b]	353.2[c]	337[c]
2015	590.9[a]	612[a]	169.7[b]	346[b]	359.8[c]	468[c]
Peer-to-peer business lending						
2016	1,350.0[a]	N.A.[a]	372.3[b]	6,244[b]	1,600.0[c]	12,968[c]
2015	2,555.0[a]	N.A.[a]	226.3[b]	3,661[b]	1,293.7[c]	11,550[c]
Angel investing						
2016	21,300.0[d]	64,380[d]	5,692.0[e]	30,230[e]	980.0[e]	8,000[e]
2015	24,600.0[d]	71,110[d]	5,109.0[e]	27,270[e]	960.0[e]	5,670[e]
Venture capital (seed stage)						
2016	2,322[f]	1,698[f]	425.5[g]	767[g]	703.9[h]	569[h]
2015	2,277[f]	1,961[f]	106.7[g]	456[g]	552.7[h]	625[h]
Venture capital (venture stage)						
2016	31,720.0[f]	2,346[f]	3,191.5[g]	2,282[g]	1,161.0[h]	372[h]
2015	35,851.0[f]	2,595[f]	2,988.3[g]	2,519[g]	1,823.3[h]	503[h]
Venture capital (growth stage)						
2016	5,964.0[f]	1,224[f]	8,617.0[g]	1,702[g]	2,597.4[h]	207[h]
2015	6,642.0[f]	1,230[f]	5,229.5[g]	1,808[g]	2,966.5[h]	265[h]

Notes: GBP converted to US$ using an average exchange rate of 0.770 (2016) and 0.681 (2015). Euro converted to US$ using an average exchange rate of 0.940 (2016) and 0.937 (2015).
Reward CF is defined as including CF projects that provide backers with non-monetary rewards, some of which may be significant enough so as the expense not to be considered a pure donation. Equity CF is defined as including CF projects that provide backers with a monetary return tied to the project's performance; includes platforms targeting accredited investors only, and those that cover the wider public. Angel investing volumes are estimates that include both the visible and non-visible market.

Sources: (a) Ziegler *et al.* (2017), (b) Ziegler *et al.* (2018),(c) Zhang *et al.* (2017), (d) Sohl (2017), (e) EBAN (2016) (f) PwC/CB Insights (2018) (*Seed* defined as containing all financing before Series A. *Venture* contains Series A through C. *Growth* contains Series D and later financing) (g) Invest Europe (2017) (*Seed* denfied as before the firm starts production for research, design. *Venture* includes funding to start, increase, or expand mass production. *Growth* is investment in a relatively mature company looking to expand or improve operations) (h) British Business Bank (2017) (*Seed* is defined as investment in companies being set up or seeking finance to establish or develop their products further. *Venture* investment is investment in companies with some years of history and in the process of gaining significant traction. *Growth* finance is investment in companies that have been alive for at least 5 years and are likely to be seeking finance to grow their core market or expand).

and France (EUR43 million) are the top three markets, whereas for reward CF, France (EUR48 million), Germany (EUR32 million) and Italy (UR34 million) are the top jurisdictions (Ziegler *et al.,* 2018).

In the US, reward and equity CF each represent about 1.6% of total seed and venture VC funding and 5% of all investments by both angel investors and VCs at the seed stage in year 2016, although it is important to note that the equity CF figure includes both equity CF proper and equity investing by accredited investors through online platforms (which, as previously discussed, could be considered a form of online angel investing rather than CF). The comparison of the US, where reward CF raised more money than equity CF, and the UK is interesting in that it illustrates the impact of regulation, much more lenient towards equity CF in the UK until year 2016, in the development of CF.

Beyond the size of the CF sector, it is important to understand the types of projects that CF finances. As Table 4 shows, projects financed through reward CF are on average substantially smaller than projects financed by any other source, although some projects financed by reward CF can be very large (*Pebble Time,* which is the project that has raised the largest amount in Kickstarter, raised over US$20 million in 2015). One can get a more complete picture of the distribution of project sizes by looking at individual platforms. At Kickstarter, the largest US reward CF platform, the median successful project in the period from 2009 to 2015 raised US$1,496, and the first and third quartiles were US$120 and US$5,796, respectively.[5] In contrast, projects financed by equity CF have an average size between those financed by angel investors and seed-stage VC deals. Although seed-stage VC deals are on average larger, the difference between equity CF deals and seed-stage VC deals is not substantial.

In terms of sectors, reward-based CF finances mainly technology, arts, and media projects in the US, continental Europe and China, whereas in the UK it also finances business services and social enterprises. This is in line with Kickstarter data, where the highest funded sectors were games (22%), design (21%), technology (20%), film (11%) and video (6%).[6] Equity CF is used chiefly to support technology, finance, real estate and internet projects in continental Europe, the UK and the US (Ziegler *et al.,* 2017; Zhang *et al.,* 2017; Ziegler *et al.,* 2018; Garvey *et al.,* 2017).

[5] Data from 15,000 Kickstarter campaigns in the period 2009–2015. The information was obtained from https://rpubs.com/dansc/kick on April 10, 2018. The data at RPubs is gathered by querying the (undocumented) Kickstarter API.

[6] Kickstarter data from https://www.kickstarter.com/help/stats accessed on April 14, 2018.

TABLE 4

DEAL SIZE BY SOURCE OF NEW VENTURE FINANCE
(Average deal size in US$ by source of new venture finance for selected regions)

	US	*EU ex UK*	*UK*
Reward-based crowdfunding			
2016	25,000[a]	16,031[b]	15,325[c]
2015	38,003[a]	4,553[b]	N.A.[c]
Equity-based crowdfunding			
2016	861,852[a]	321,937[b]	1,048,330[c]
2015	965,361[a]	489,864[b]	769,424[c]
Peer-to-peer business lending			
2016	N.A.[a]	118,759[b]	123,377[c]
2015	85,902[a]	106,708[b]	112,012[c]
Angel investing			
2016	330,185[d]	188,290[e]	112,500[e]
2015	345,390[d]	187,349[e]	169,312[e]
Venture capital (seed stage)			
2016	1,367,720[f]	554,800[g]	1,558,442[h]
2015	1,161,270[f]	234,043[g]	1,152,717[h]
Venture capital (venture stage)			
2016	13,520,827[f]	1,398,549[g]	3,896,104[h]
2015	13,815,449[f]	1,186,288[g]	4,111,601[h]
Venture capital (growth stage)			
2016	22,319,395[f]	5,062,880[g]	15,844,156[h]
2015	31,416,845[f]	2,892,398[g]	14,390,602[h]

Notes: GBP converted to US$ using an average exchange rate of 0.770 (2016) and 0.681 (2015). Euro converted to US$ using an average exchange rate of 0.940 (2016) and 0.937 (2015).
Reward CF is defined as including CF projects that provide backers with non-monetary rewards, some of which may be significant enough so as the expense not to be considered a pure donation. Equity CF is defined as including CF projects that provide backers with a monetary return tied to the project's performance; includes platforms targeting accredited investors only, and those that cover the wider public. Angel investing volumes are estimates that include both the visible and non-visible market.

Sources: (a) Ziegler *et al.* (2017), (b) Ziegler *et al.* (2018), (c) Zhang *et al.* (2017), (d) Sohl (2017), (e) EBAN (2016), (f) PwC/CB Insights (2018) (*Seed* defined as containing all financing before Series A. *Venture* contains Series A through C. *Growth* contains Series D and later financing), (g) Invest Europe (2017) (*Seed* defined as before the firm starts production for research, design. *Venture* includes funding to start, increase, or expand mass production. *Growth* is investment in a relatively mature company looking to expand or improve operations), (h) British Business Bank (2017) (*Seed* is defined as investment in companies being set up or seeking finance to establish or develop their products further. *Venture* investment is investment in companies with some years of history and in the process of gaining significant traction. *Growth* finance is investment in companies that have been alive for at least 5 years and are likely to be seeking finance to grow their core market or expand).

III. THE THEORY OF CROWDFUNDING

The growth of CF has triggered a rapid increase in the number of theoretical papers that try to capture the motivation for and the implications of the use of CF. Although very few of these papers have been published at the time of writing this review, we have opted to be quite comprehensive in our review of the theoretical literature, to help researchers identify the questions that have received an adequate treatment so far, the main gaps that remain, and the elements that emerge from this first round of models as key to understand CF. As we mention in the introduction, we restrict our attention mainly to papers that, methodologically, lie in the fields of economics and finance.

Most of the analysis of the financing of entrepreneurial ventures focuses on the asymmetries of information between the entrepreneur, who observes the quality of her project, and the investor, who may not observe the project's quality or the entrepreneur's ability. However, the feature of CF that has received the most attention is precisely that it may allow the entrepreneur to learn about the value of the project and, thus, invest only if the expected value is high enough. In other words, whereas traditional screening mechanisms allow investors to access information initially possessed by the entrepreneur, without generating new information about the project's value, the screening performed by CF generates new information about project value by aggregating pieces of information dispersed among potential backers. As we will see in greater detail below, CF can therefore play the role of an incentive compatible market research tool. As such, entrepreneurs may find it optimal to use CF even if they actually do not need any funding.

Although most models of CF focus on the screening role of CF as a mechanism to "harness the wisdom of the crowd," several models also analyze how this screening role is affected by traditional problems in corporate finance, such as entrepreneurial moral hazard or private information about project quality possessed by the entrepreneur.

To unify the treatment of the different models and better understand their commonalities and differences, consider the following general framework. The entrepreneur (E) may invest in a project to produce a new good at a fixed cost c. There is a set $N = \{1, ..., n\}$ of potential consumers, who may buy either a unit of the good or nothing, and a set $M \subseteq \{1, ..., n + m\}$ of potential backers who may participate in the CF campaign. The sets N and M may coincide, may have only some elements in common (for example, one may be a subset of the other), or may have an empty intersection. Each potential consumer i has a characteristic θ_i and receives a signal $\hat{\theta}_i$, with θ and $\hat{\theta}$ being the corresponding vectors of characteristics and signals. Each backer j (who may also be a consumer

if $j = i$ for some $i \in N$) receives a signal $\hat{v}_j$. Finally, E has some characteristic ω, receives a signal $\hat{\omega}$, and may or may not take some action a after the CF campaign. The general interpretation of the consumer's characteristic θ_i is that it is either i's true valuation of the good or a parameter that determines, together with other variables, i's valuation of the good or its distribution. The signals $\hat{v}$ are usually signals about the value of the good for consumers, or, more generally, about the distribution of that value. E's characteristic ω may capture the quality of the good or some parameter that determines either the production function of the good or the probability distribution of the quality of the good, and $\hat{\omega}$ is a signal of ω. Finally, the action a generally captures some action that E can take to influence the quality of the good. This general framework encompasses a large variety of scenarios, and most existing models can be described as particular formulations. The models differ in the definition of N and M, in the joint distribution of $(\theta, \hat{\theta}, v, \hat{v}, \omega, \hat{\omega})$, and in the variables that each player observes.

We first review the more numerous models of reward CF (RCF) and then turn to the models that analyze equity CF (ECF). Throughout our discussion we employ the pronoun *she* to refer to the entrepreneur and *he* to refer to a consumer or investor.

1. Models of Reward Crowdfunding

1.1. Benchmark RCF Model of Learning About Demand

In the simplest possible model of RCF as a mechanism to learn the value of E's project, each consumer has a valuation of the good, which can be either 0 or 1 (*i.e.*, $\theta_i \in \{0, 1\}$), and consumers' valuations are *i.i.d.*[7] Therefore, for any price not greater than one, the demand for the good would be just the number of consumers with valuation 1 (*i.e.*, $\sum_{i=1}^{n} \theta_i$) and the total surplus generated by E's project if all consumers obtained the good would be $\sum_{i=1}^{n} \theta_i$. Each consumer observes only his own valuation, and E cannot observe any of the valuations and only knows the distribution of θ_i.

In this simple model, there is demand uncertainty, since the number of consumers who value the good is uncertain. Moreover, the information about demand is dispersed among consumers, each observing only his own valuation. Since E has to sink a cost c to produce the good, E would like to know whether demand is high enough to justify the investment.

[7] Strausz (2017) provides a similar motivating example.

In this setting, an RCF campaign in which the potential backers are the consumers (*i.e.,* in which $M = N$ and $\hat{v}_i = \hat{\theta}_i = \theta_i$) may allow *E* to sink the fixed cost *c* and produce the good if and only if demand is higher than the fixed cost of production. To achieve this goal, *E* can run an AoN campaign in which: (i) backers may make a *pledge p,* (ii) production takes place if and only if the total amount pledged by backers is at least equal to some *funding threshold T,* and (iii) the good is sold only to backers (or the price of the good if there is a post-campaign market is equal to 1). Since consumers pay *p* only if the good is produced, it is optimal for them to pledge if and only if they value the good as long as $p \leq 1$. Therefore, by setting the pledge equal to 1 and the funding threshold *T* equal to the fixed cost *c* of producing the good, *E* achieves the goal of producing the good if and only if the total surplus is nonnegative and, moreover, is able to extract all the surplus. Thus, in this very simplified scenario, an AoN RCF campaign is not only clearly superior to the alternative of first deciding whether to sink the fixed cost *c,* but it is actually optimal for *E.*[8]

In the above paragraph, we purposefully omitted any reference to the funding of *E*'s project. The optimal CF design would allow *E* to fund the project whenever the campaign threshold is met, so that *E* would be able to finance the project via CF. However, even if *E* had *c* available to finance the project herself, she would still find it optimal to make her production decision and sell her product using the optimal CF design. In other words, the *funding* in *crowdfunding* is not necessary to make the mechanism just described attractive for entrepreneurs. In fact, *E*'s problem is like the problem of a monopolist facing uncertain demand (see, *e.g.,* Cornelli, 1996) or the problem of a planner who has to decide whether to provide an excludable public good, problems that had been studied in economics extensively before the advent of CF (see, *e.g.,* Palfrey and Rosenthal, 1984).[9]

This example showcases the role of CF as a mechanism for entrepreneurs to learn about and adapt to the potential demand for their products by aggregating the dispersed information possessed by potential consumers. Surely, entrepreneurs could attempt to obtain that information from surveys or by applying other market research techniques, but conducting a survey may be very costly and has the additional problem that it may be difficult to convince potential buyers to both participate (since participation does not determine the possibility of enjoying the good) and reveal their true valuations (since exaggerating one's valuation has no cost).

[8] In more complex environments with more than one type of consumer with a positive valuation for the good, a standard AoN campaign need not be optimal.

[9] In these literatures, a simple mechanism in which the provision of the public good takes place only if a threshold level of total contributions is met is called a *provision* point mechanism.

Many models of reward CF provide variations and extensions of this simple example. Before describing these models, we first describe another building block of many of the models that capture the ability of CF to obtain information from consumers.

1.2. Benchmark Price Discrimination Model without Uncertainty

Consider now a case in which there is a greater range of valuations among those consumers who derive a positive value from consuming the good. For example, assume that $\theta_i \in \{\theta_L, \theta_H\}$ with $0 < \theta_L < \theta_H$. As above, suppose that each consumer observes his type. However, suppose now that the number of consumers with high valuation and the number of consumers with low valuation are known, so that there is no uncertainty about demand, but *E* cannot observe which consumers have high valuation and which consumers have low valuation.

Suppose first that *E* does not run a RCF campaign and instead, decides first whether to sink the fixed cost *c* and then the price for her good. Recalling that *n* is the number of consumers, if we assume that $\sum_{i=1}^{n} \theta_i \geq c$ then *E* knows that there is enough demand to cover the fixed cost of production. If *E* could observe each consumer's valuation, she would be able to charge θ_L to low-valuation consumers and θ_H to high-valuation consumers, so that not only would the good be efficiently produced, but *E* would extract all the surplus. However, since *E* cannot observe each individual's valuation, if she produces, she will charge either θ_H (if there are enough high-valuation consumers) or θ_L (otherwise). If $n\theta_L \geq c$ or $\sum_{\theta_i=\theta_H} \theta_i \geq c$, *E* will pay the fixed cost *c* and sell the good at a price equal to θ_L or θ_H depending on the number of consumers of each type. However, if neither of these conditions holds, *E* will, inefficiently, not produce the good even if $\sum_{i=1}^{n} \theta_i \geq c$.

E may improve on this outcome by means of an AoN RCF campaign. In particular, if *E* allows for two possible pledge levels, equal to θ_H and θ_L, and sets the funding threshold equal to $T = \sum_{i=1}^{n} \theta_i \geq c$ there is a Nash equilibrium in which all consumers pledge their true valuation. To see why such a Nash equilibrium may emerge, consider a high-valuation consumer *j*. If *j* expects others to pledge according to their valuation, he expects the total amount pledged by others to be $\sum_{i \neq j}^{n} \theta_i$. Since the funding threshold *T* is equal to $\sum_{i=1}^{n} \theta_i = \sum_{i \neq j}^{n} \theta_{i-} + \theta_j$, if *j* pledges his true valuation, the good is produced, but if be does not pledge any money or makes the lower pledge θ_L, the good is not produced. In other words, consumer *j* is *pivotal* in that whether the project is carried out depends on his pledge. The possibility of being pivotal generates incentives for backers to contribute, instead of free-riding on others' contributions. Moreover, the fact

that the pledges are refunded if the threshold is not achieved reduces the risk of pledging.

As in the case in which *E* uses CF to learn about the level of demand, this example shows that *E* may want to use RCF as a tool for price discrimination, even if she does not need any financing. An important implication of this example (discussed by Kumar *et al.,* 2016 and Ellman and Hurkens, 2016) is that, even though such price discrimination may reduce consumer surplus in some contexts, it may increase total surplus without reducing consumer surplus in contexts in which production is efficient but would be unfeasible with traditional selling.

Kumar, Langberg and Zvilichovsky (2016) propose a model very close to this example, except that price discrimination does not take place by means of several pledge levels at the CF stage, but by charging different prices to the consumers who contribute to the campaign (who pay a price equal to the pledge level), and the consumers who buy on the post-campaign spot market (who pay a price equal to the monopoly price set ex post by *E*). By using an AoN design, *E* can make some consumers pivotal and thus induce them to pay a pledge higher than the expected spot market price, because their pivotality implies that the alternative to not pledging is not enjoying the good at the post-investment price, but, rather, not enjoying the good at all. Sayedi and Baghaie (2017) make this point as well in a model that also considers signaling on *E'*s side. Although CF may be used by a financially unconstrained *E,* Kumar, Langberg and Zvilichovsky (2016)) also analyze the case in which external finance is costly. In particular, they study how the cost of external finance determines the optimal pledge and funding threshold and whether *E* uses RCF. Kumar, Langberg and Zvilichovsky (2016) show that *E* will use RCF exclusively when she has enough wealth to finance the investment (so external funds are not required) or when the cost of external finance is sufficiently high, and will otherwise use both RCF and external finance. Kumar, Langberg and Zvilichovsky (2016) also show that a greater need for external finance or a higher cost of external finance will lead *E* to set lower pledge levels and higher funding thresholds to ensure that RCF contributes more to the financing of production. An interesting feature of this model is that it relates the use of CF and the design of CF campaigns to standard corporate finance variables, such as the availability of internal funds and the cost of external finance. However, external finance is treated in a very reduced-form manner, so that, for example, the required interest rate is independent of the design and outcome of the CF campaign.

Belleflamme, Lambert and Schwienbacher (2014) also provide a model in which RCF allows *E* to price-discriminate and charge higher prices, in the

form of campaign pledges, to those consumers with a higher willingness to pay for the good. Belleflamme, Lambert and Schwienbacher (2014) assume that consumers also derive utility from participating in a CF campaign, but that such utility is different in RCF and ECF. Participating in an RCF campaign allows consumers to have an impact on the design of the good and has a greater impact on the utility of those consumers who value the good more. In contrast, participation in ECF provides the benefit of making the investment happen, which is assumed to be independent of consumers' valuations. Belleflamme, Lambert and Schwienbacher (2014) exploit this difference to explain which form of CF is optimal in different contexts.

The model by Ellman and Hurkens (2016) combines the two benchmark models of demand learning and price discrimination to derive implications about the optimal funding threshold and pledge levels in an RCF campaign. Before discussing their contribution, it is important to note that the industrial organization and mechanism design literatures had already addressed similar problems and, in particular, Cornelli (1996) had already proposed an optimal indirect mechanism in a setting very similar to the one studied by Ellman and Hurkens (2016) and most of the other models on the adaptation benefits of RCF. Ellman and Hurkens (2016) contribute to the existing mechanism design literature by showing that a standard RCF design is optimal only in the very particular setting with two types of consumers who value the product. However, the main contributions of their paper to our understanding of RCF arguably lie elsewhere. First, they show how the optimal RCF campaign (as well as the resulting production decisions, profits and welfare) depends on the investment size and the *ex ante* probability that the demand for the product is high. Second, they consider several extensions of their model to analyze, among other things, how the motivations of entrepreneurs (who may be purely profit-motivated or care instead about project success or total welfare) affect optimal RCF design and outcomes, the determinants of the self-selection of entrepreneurs with different types of projects into RCF, the value of RCF to predict not only current demand but also future demand, and the interaction of RCF with traditional financing. Importantly, they also study the effect of crowd size on the value of RCF. This is a crucial issue, since a large crowd, by lowering the probability that any investor becomes pivotal, weakens the ability of RCF to charge high prices to high-valuation consumers. However, Ellman and Hurkens (2016) show that the benefits of RCF survive even for relatively large crowds.

The models considered so far analyze the screening value of RCF. However, in contrast to traditional models of early stage financing, in which the key goal of the design of a financial vehicle is to allow providers of funds to screen entrepreneurs or projects, the goal of RCF in the models discussed so far is to allow the entrepreneur to screen consumers who may possess different valuations

for the good. This shift of focus and the fact that in these models there is no uncertainty about the provision of the good by the entrepreneur if the funding threshold is met imply that these RCF models could be equally considered models of pre-sales. Although this role of RCF as a pre-sales mechanism is likely to be very important, one could argue that there are other essential elements of reward CF that set it apart from pure pre-sales mechanisms and that the pure screening models of the type just discussed lack.

1.3. Entrepreneurial Moral Hazard

A key element of many models of early stage financing is the existence of a moral hazard problem on the side of the entrepreneur. Once the entrepreneur has obtained financing for her project, she may not use it in the ways that maximize the return of investors and may even abscond with the money, either literally or by claiming that the investment could not be finalized for reasons beyond her control. This extreme possibility is especially relevant in the case of RCF, since the amount of due diligence performed by funders is likely to be small and since, as emphasized by Gutiérrez Urtiaga and Lacave (2018), standard RCF contracts offer close to no protection to backers in case they do not receive their rewards. Therefore, several models consider how moral hazard may affect the outcomes of RCF and whether a correct design of the RCF campaign may also help control managerial moral hazard. Strausz (2017), Chemla and Tinn (2017), and Chang (2016) all consider different versions of the learning model described in Section III.1.1 and incorporate the possibility that the entrepreneur may run away with the money.

Strausz (2017) shows that the optimal mechanism involves both deferred payments (*i.e.,* some of the money pledged is not given to *E* until after the rewards are distributed) and no revelation of the amount of overfunding (*i.e.,* the amount pledged above the funding threshold) if the campaign is successful, and argues that RCF campaigns can be interpreted as having these two features because there is a post-campaign market, and because the dynamic nature of usual RCF campaigns implies that many backers may refrain from contributing once the target has been met.

Chemla and Tinn (2017) and Chang (2016) explicitly show that the existence of a post-campaign market may deter *E* from absconding with the money if the expected profitability of that post-campaign market is high enough. The interesting feature of their models is that they show how the design of the CF campaign may increase the expected profitability of that market conditionally on the pledge threshold being met. Their models thus show how the determination

of the pledge level and the funding threshold should balance the objectives of learning consumers' valuations, so as to invest when it is profitable, and to provide incentives for *E* to carry out the investment, so that backers are willing to contribute. In particular, both papers argue that, since the incentives to carry out the investment will be stronger when expected after-campaign demand is higher, *E* should receive the money when expected demand is high and not when expected demand is low. Since the demand by backers is predictive of the demand in the after-campaign market, this goal can be achieved by setting a higher funding threshold. By doing so, the threshold will be met only in cases in which the demand by backers is sufficiently high.

1.4. Signaling through Reward Crowdfunding

The emphasis in all the models consider so far is on the ability of RCF to aggregate information about the value of *E*'s project that is dispersed among consumers. However, even if that kind of information is undoubtedly of first order importance, the information about *E*'s competence or trustworthiness, the technical feasibility of the project, or the true quality of the project is also likely to be important. A relevant question is thus how CF may help consumers or investors learn this kind of information, which, in many cases, one expects *E* to possess.

Chakraborty and Swinney (2017) consider a setting in which consumers' taste for quality is known but not the total number of consumers and in which the uncertain value of the good for consumers depend on its quality. Chakraborty and Swinney also assume that some consumers (uninformed consumers) cannot observe the quality of the good, whereas others (informed consumers) can observe the good's quality. In this setting, high-quality Es may signal their type to consumers by setting a high funding threshold. Low-quality *E*s may not imitate, because the presence of informed consumers implies that, even if uninformed consumers decide to pledge thinking that *E* is of high quality, the funding threshold may not be met because informed consumers will not pledge. An interesting implication of the model by Chakraborty and Swinney (2017) is that the quality of the projects that attempt to obtain financing via RCF will depend on the expected fraction of informed consumers. Thus, when the fraction of informed consumers is very small, high-quality *E*s will have to increase their funding threshold excessively to achieve separation, so that the benefits from the RCF campaign will be reduced. The lower profitability of the RCF for high-quality *E*s, in turn, may lead these Es to shun RCF (or to opt for low quality projects *ex ante*). Chakraborty and Swinney (2017) also consider how platforms should set their fees considering their impact on the selection of project quality.

Sayedi and Baghaie (2017) consider a different setting, in which there is a known distribution of consumer valuations for product quality conditional on the level of aggregate demand, which is uncertain, but in which *E* does not observe consumer valuations for quality. In this setting, Sayedi and Baghaie (2017) assume that project quality is decided by *E* after the campaign takes place and higher quality is more costly to produce. Therefore, the moral hazard problem is not that *E* may run with the money but that she may produce a good of low quality. At the same time, the cost of quality depends on the *E*'s competence, which is privately observed by *E*. The contribution of Sayedi and Baghaie (2017) is to show how the design of the RCF should optimally balance the goals of price discrimination, of providing incentives to *E* to select high quality if the project is funded, and of signaling competence. As in the models of Chemla and Tinn (2017) and Chang (2016), the incentives to behave well if the project is funded stem from the existence of an after-campaign market in which the quality of the good becomes known. However, Sayedi and Baghaie (2017) assume that the size of the after-campaign market is endogenously determined as the difference between the size of the market and the size of the fraction of consumers who decide to pledge. This assumption implies that, contrary to the models by Chemla and Tinn (2017) and Chang (2016) (which effectively force after-campaign demand to be increasing in the value of the project), a higher funding threshold reduces the size of the after-campaign market. Therefore, high-competence producers set a high price and a low funding threshold to separate from low-competence producers. Such a policy implies that very few consumers participate in RCF, and that, as a consequence, the after-campaign market is large and, thus, the benefits of setting a high quality are large. Since the cost of quality is larger for the low-competence *E*s this strategy is more costly for them, which allows high-competence *E*s to separate. As in the model by Chakraborty and Swinney (2017), if signaling requires a large distortion in the pledge level and funding threshold (as it may be the case when the difference between competent and incompetent producers is large), then competent producers would obtain little gains from price discrimination if they used RCF, and may thus opt out of RCF.

2. Models of Investment Crowdfunding

2.1. Simultaneous Contributions

Brown and Davies (2018) study a common value setting to analyze the ability of equity CF (ECF) to aggregate the information possessed by dispersed investors, who are not potential consumers of the good. In the model, E's project has a fixed size and ECF consists of an AoN campaign in which contributing investors receive a share of the firm's realized value proportional to

their contribution if the sum of the contributions reaches the funding threshold. Brown and Davies highlight two sorts of distortions that may lead investors not to use the information contained in their signals. The first sort of distortion, which they label the *loser's blessing,* is due to the fact that the funding threshold implies that if enough other investors are acting on their private information, then, if the project is bad, it is likely that it will not be carried out and, thus, that contributions will be returned. This hedging of the risk that the project is bad leads investors to contribute more aggressively, *i.e.,* to contribute even if they receive negative information. On the other hand, the fact that the project is of fixed size implies that, if the project is good, many investors are likely to participate and, thus, the fraction of the project's value accruing to each contributing investor is small. If the project is bad, however, fewer investors are likely to participate, so that each of them will have a claim to a relatively large share of the firm's low value. This winner's curse gives incentives to investors to be more conservative in their bids. Brown and Davies (2018) show that either because of the loser's blessing or the winner's curse, the ECF campaign never aggregates optimally the information dispersed among investors. Moreover, with a continuum of investors, the ECF campaign would actually not be able to extract any information from investors, who would either invest or not invest regardless of their signal.[10]

Hakenes and Schlegel (2014) provide a model of AoN crowdinvesting in which *E* offers debt securities, instead of equity, to investors, who are not potential consumers of the good, to finance a risky project. If they exert costly effort, investors receive independent signals (conditionally on the value of the project) of the probability of success of the project. Hakenes and Schlegel focus on the case in which *E* knows the project's probability of success, so that she does not benefit from the wisdom of the crowd (although they also briefly analyze the case in which she does not know the project's success probability). Hakenes and Schlegel also assume that the project has negative NPV if its probability of success is low, which guarantees that there can be no separating equilibria in which *E*s with high and low probability of success offer different combinations of interest rate and funding threshold. Since the choice of campaign cannot signal the quality of the project in equilibrium, the *E*s with high-quality projects would like to design the campaign in such a way that investors have an incentive to exert effort and to pledge if and only if they get a good signal. Indeed, Hakenes and Schlegel show that, in equilibrium, *E*s set the interest rate and the funding threshold in such a way that all investors become informed. Moreover, they show that in equilibrium there is too much information acquisition and that whether CF increases welfare relative to standard debt financing depends on parameter values.

[10] Gruener and Siemroth (2017) also analyze an ECF model and focus on the effect that the distribution of wealth across consumers has on the outcomes implemented through ECF.

2.2. Dynamic Contributions and Herding

In all the models considered so far, backers are assumed to make their contributions simultaneously, without having obtained any information about other backers' signals or contributions. However, in typical CF campaigns, individual pledges are made public as soon as they are made (in some cases, even the identity of the backer behind each pledge is public), so potential backers can condition their pledge decisions on past pledges. This scenario can lead to herding behavior by investors, which may result in a failure to incorporate the private information obtained by each investor.

Herding is said to occur when individuals disregard the private signals they obtain about the value of an asset and base their investment decision only on their observation of the actions of others (Banerjee, 1992; Welch, 1992; Bikhchandani, Hirshleifer and Welch, 1998). Herding may lead to informational cascades if investors disregard of their own signals persists over time. Importantly, although herding and cascades may be the result of irrationally imitative behavior, herding may also be rational in the sense that the decision to disregard one's signal follows from optimal Bayesian updating upon the observation of prior investors' actions.

So far, two papers have analyzed the possibility of rational herding and informational cascades in equity CF, reaching different conclusions. Astebro *et al.* (2018) show that rational herding can occur in equilibrium, leading investors to either invest when they receive a bad private signal, or to abstain from investing when they receive a good signal, but only the latter kind of herding can lead to a cascade that causes the failure of the campaign.

Cong and Xiao (2018) propose a model that essentially adds an AoN provision to the classic model of Welch (1992) and obtain a result opposite to that of Astebro *et al.* (2018). In particular, Cong and Xiao find that if the funding target is set optimally by the sponsor, in equilibrium there can be *up* cascades (in which all investors invest regardless of their private information from some point onwards) but no *down* cascades.

The contrasting results of the two papers point at the need for a careful selection of the assumptions that best characterize equity CF campaigns (and a careful discussion of the generalizability of the results), as well as to the possibility that the details of CF campaign design may have a substantial effect on equilibrium campaign paths, which would be an important implication for platforms, issuers and regulators. Apart from the aspects differentially considered by Astebro *et al.* (2018) and Cong and Xiao (2018– such as whether the contribution of each backer is fixed–other potential features of CF campaigns, such as the possibility

that investors derive utility from the success of the campaign (as consumers, as family or friends, or as "community") and that this utility is heterogeneous across investors (which means that contribution decisions may respond to differences in private valuations and not only to differences in signals about the project common value), or the lack of sophistication by some contributors may lead to dynamic effects absent in other contexts.

3. Contracts

Most of the theoretical research on VC aims at explaining the numerous contractual provisions incorporated in the securities used by VCs to finance entrepreneurial firms. These provisions often have a large degree of conditionality (by means of convertibility or explicit conditions that may trigger different actions) and have to do not only with the payoffs to security holders, but also with their control rights and those of the entrepreneur.

Given the history of theoretical research on VC, one would have expected similar attention to be focused on the design of CF contracts. However, with the exception of the paper by Gutiérrez Urtiaga and Lacave (2018), there have been, to our knowledge, no attempts to model CF contracts, reward or equity, beyond the choice of pledge level and funding threshold. In our view, this is a major gap in the theoretical analysis of CF.

Gutiérrez Urtiaga and Lacave focus on the penalties for non-delivery of the good in RCF. They highlight that current RCF contracts contain essentially no penalty for non-delivery, since the entrepreneur has no liability as long as she conducts "best efforts" to deliver the good. Although a penalty could help mitigate moral hazard on the side of the entrepreneur, Gutiérrez Urtiaga and Lacave show that, in some contexts, doing so is not optimal. They consider a model in which the entrepreneur may decide not to deliver the good after a successful campaign, so as to avoid incurring the delivery costs. Two motivations may deter the entrepreneur from doing so. The first one is the penalty. The second one the fact that, in case the entrepreneur's ability is sufficiently high, she can obtain high expected profits in the post-campaign market, say, after receiving VC funding. The decision to deliver the good may act, thus, as a signal that the entrepreneur's ability is high, which increases the entrepreneur's expected post-campaign profits. However, the value of the signal is reduced if the penalty can already induce entrepreneurs with both high and low ability to deliver the good to backers. Gutiérrez Urtiaga and Lacave show that this effect of the penalty implies that in contexts in which both the level of ability required to benefit from scaling up and the benefits from scaling up are large, a zero penalty may be optimal.

4. The Relation with VC and other Financing Sources

A prominent feature of early stage financing is that initial rounds of financing are often followed, if the project is successful, by new rounds of financing, so that financing is, explicitly or implicitly, staged. Several models consider the impact of post-CF financing rounds on the design and outcomes of CF. For example, Strausz (2017) argues that the possibility of obtaining future VC financing may provide the deferral of payoffs necessary to induce the entrepreneur not to abscond with the money raised with RCF. Ellman and Hurkens (2016) show that RCF can reduce demand for other forms of financing in some contexts, but it may also increase the use of these other financing sources in others, either by allowing (joint) financing in contexts in which financing would not have occurred in the absence of RCF, or by generating information about future demand that makes the project attractive for the providers of these other sorts of funding. Kumar, Langberg and Zvilichovsky (2016); Chen, Gal-Or and Roma (2017); Babich, Tsoukulas and Marinesi (2017), or Schwienbacher (2017) also consider the interaction between RCF and other financing sources. However, although these papers highlight some interesting trade-offs, the analysis is still of a preliminary nature, since either CF or the alternative financing source are modeled in a very reduced form way. For example, the interest rate required by investors after a CF campaign is assumed to be independent of the design and outcome of the campaign (Kumar, Langberg and Zvilichovsky, 2016) or, alternatively, the CF campaign is assumed to be a black box that generates a fixed amount of money and a fixed signal if successful (Babich, Tsoukulas and Marinesi, 2017).

5. Platform Design

Although an entrepreneur could, in principle, carry out a CF campaign on her own, CF platforms (CFPs) greatly reduce the transaction costs of organizing a campaign and offer entrepreneurs a wider reach. CFPs can constrain, and typically do, the space of possible campaign designs that entrepreneurs can offer through them. For example, some platforms (*e.g.,* Kickstarter) require campaigns to be AoN, whereas others allow for KIA designs as well (*e.g.,* Indiegogo). Platforms could decide whether backers bids are submitted simultaneously or sequentially and, in the latter case, they can also decide the duration of the campaign and what information is provided on the platform as the campaign progresses. Platforms can likewise establish selection criteria and perform no or different levels of due diligence before a campaign starts, and may share with potential backers some of the information gathered during the due diligence process. Platforms could also help backers pick campaigns and act as intermediaries with VC firms or angel investors. Belleflamme, Omrani and Peitz (2015) discuss several of the design dimensions of CFPs.

Since platforms can have such a large impact on the design and outcomes of CF campaigns, the research on CF should shed light on the incentives of CFPs and the effect of those incentives on CFP design and, in turn, on the design and effectiveness of the CF campaigns conducted through them. In particular, it seems to us that very little can be said about the regulation of CF without understanding how CFPs decide the structure of CF campaigns and how regulation may affect their choices. For example, even if one shows that some kind of CF campaign is optimal in a certain context, CFPs may not have the incentives to allow or promote such a design. However, despite the essential role of platforms in CF, very little research so far has addressed in a meaningful way the design of CFPs and the incentives of the managers of CFPs. As mentioned above, some papers do consider how the fee charged by platforms to entrepreneurs may affect the latter's choice of campaign parameters and how platforms may take that into account when determining fees (*e.g.,* Ellman and Hurkens, 2016). Other papers dwell on the decision whether to offer AoN or KIA campaigns, but, overall, the analysis has been extremely limited.

A key feature of CFPs, emphasized by Belleflamme, Omrani y Peitz (2015), is that they are two-sided platforms that bring together backers, on one side, and entrepreneurs, on the other. A sizeable literature in industrial organization has analyzed the incentives of two-sided platforms and the effects of platform competition on, among other things, pricing strategies or the provision of information to either side of the market.[11] Belleflamme, Omrani y Peitz (2015) argue that there are positive within-side network effects among backers and positive cross-side network effects, but direct within-side network effects are negative for entrepreneurs and, in the light of the literature on two-sided platforms, comment on the possible implications of such network effects for CFPs' pricing strategies (such as offering subscription fees, fees per campaign, proportional fees conditional on campaign success, and so on).

The implications of the fact that CFPs are two-sided platforms for the form in which they structure CF campaigns is an open question. The potential role of CFPs as certification intermediaries and the impact of the competition between platforms on this role are also issues that would benefit from a careful theoretical analysis.

From the point of view of regulation, it is essential to understand the CFP market: Will CFPs structure campaigns in an efficient way? If not, what are the main frictions that preclude them from doing so? Can regulation help? What is the role of CFP competition in determining the efficiency of the CFPs' campaign structure, pricing, and services? Is platform consolidation a good or a bad thing? Should platforms be regulated as financial intermediaries? Should

[11] See Rochet and Tirole (2006) or Armstrong (2006) for reviews of the literature on twosided markets.

one welcome or try to prevent the integration of CFPs with banks or investment advisers?

6. Post Financing Monitoring

A key role of angel investors and VCs is advising and monitoring the managers of the companies they finance, and several models of VC financing focus on the implications for contract design and outcomes of investors' monitoring effort (see, *e.g.,* Casamatta, 2003; Hellmann, 2006). Individually, CF backers may not be able to exert much influence on the entrepreneurs they fund. However, they may have the motivation and knowledge to monitor and advise entrepreneurs (crowd-monitoring), and the CFP could also help with that role. However, to our knowledge, and in contrast to the VC literature, there is no theoretical model that analyzes the potential provision of monitoring and advising by CF backers and how CF campaigns and contracts may be structured to elicit such monitoring optimally.

7. Discussion and Suggestions for Future Theory Research

The above discussion of the theoretical models of CF identifies two major gaps in our theoretical understanding of CF. The first one is the lack of models that, in the spirit of the security design literature, study the optimality of different contracts between backers and the entrepreneur, and, possibly, the platform as well. The second major gap in the literature is the lack of a theory of CFPs. Filling these gaps appears to us as a very promising avenue for future research.

The theoretical literature mostly takes as given the design of CF campaigns, which it typically characterizes with just two variables (the pledge level and the funding threshold), and analyzes how different parameters may affect the optimal choice of these variables. However, there are many possible design dimensions that have been largely unexplored, such as whether contributions should be made simultaneously or sequentially, or the information that should be revealed to potential backers during a campaign. Moreover, the literatures on auctions and public good provision may suggest different mechanisms to compare to the provision point ones typically used by CFPs.

A greater conceptual integration with the literatures on VC and angel investing would also shed light on the nonmonotonic relations that seem to plague new venture financing. For example, the literature on VC argues that the large degree of uncertainty and opacity that characterize early stage ventures

requires that financing be provided by specialized intermediaries, who protect their investments by means of contracts that limit the entrepreneur's choices and confer different kinds of control rights to investors. However, the seed and very early stages, which are typically served by angel investors, are characterized by even greater uncertainty, yet the financing arrangements between angels and entrepreneurs seem to be simpler and substitute informal monitoring by the angel investor for formal contractual restrictions and control rights. The kinds of projects financed with CF are, arguably as uncertain or more than those funded by angel financing. However, again, CF replaces the informal monitoring activity of the angel investor by little or no direct monitoring by the dispersed backers and, in the case of equity CF, by contracts that often resemble the ones used in VC. Beyond this conceptual integration, much could be gained by incorporating VC, angel financing, or bank financing in a meaningful way in models of CF.

Given the widespread concern that small investors may not have the skills, the time, or the resources to process all available information optimally or may not have access to all relevant information, explicitly incorporating these limitations into theoretical models of CF would be extremely useful, especially for regulators. Although in a very reduced-form way, the paper by Hornuf and Schwienbacher (2017) is a first attempt at providing models that explicitly address the kinds of trade-offs that concern regulators.

Our review of the theoretical literature on CF showcases that, even within the narrow confines of the standard AoN campaign, there are many modeling choices open to the researcher, and apparently minor differences in the assumptions (for example, whether the investment project is scalable or not, or whether the entrepreneur may obtain financing from other sources if the campaign fails) may have important consequences for the results. It would be useful if future rounds of theory papers provided a more thorough justification of their "auxiliary" assumptions and discussed the robustness of their results to changes in those assumptions. A more careful description of the models' empirical predictions would also be very useful for empirical researchers.

IV. EMPIRICAL WORK

The focus of most empirical work on CF has been on the determinants of campaign success. In general, empirical work has been little informed by the economic theory on CF that we describe in Section III. In what follows, we provide a selective review of the empirical literature, trying to embed it in the framework presented in Section II.3 and to relate it to the theoretical work. The empirical literature on CF is already vast. In this review, we focus on

published papers in the economics, finance and management fields. We also refer to industry reports when necessary.

1. The Design of Crowdfunding Campaigns

As discussed in Section II.3, the main features of the design of a CF campaign are the funding threshold (which is zero in the case of KIA campaigns), the menu of possible contributions and corresponding rewards, and the length of the campaign. Some of the theoretical models provide comparative statics results about these features of campaigns. However, the empirical work has not tested these predictions so far, and has paid little attention to the determinants of the main features of CF campaigns. Mostly, campaign design features appear as explanatory variables or controls in regressions that predict campaign success. We now summarize the available information about the design of CF campaigns.

AoN vs KIA. Cumming, Leboeuf and Schwinenbacher (2015) report that the *median campaign target* (on Indiegogo) *is larger in AoN campaigns than in KIA campaigns* (US$16,485 vs. US$10,000) and that *AoN campaigns attract more backers than KIA campaigns* (median 43 v. 33). They find that small scalable projects are more likely to be funded through KIA campaigns and that, controlling for size and other determinants of success, *KIA campaigns are less likely than AoN campaigns to achieve their funding goals.*

Funding target. In agreement with the evidence that we document in Section II.4 in relation to the funds raised by crowdfunded projects, *funding targets for reward CF are very small.* For example, in the period from 2009 to 2015, the median funding target among Kickstarter's projects was US$5,000 (mean of US$31,170).[12] Mollick (2014) also reports (for Kickstarter) that average *funding targets differ significantly by industry.*

Several papers find that *the size of the campaign's funding goal is negatively associated with success probability for reward CF* (Mollick, 2014; Crosetto and Regner, 2014; Cumming, Leboeuf and Schwinenbacher, 2015) and for equity CF projects (Hornuf and Neuenkirch, 2017).

The *funding targets of equity CF campaigns are significantly larger than those of reward CF campaigns.* For example, Vulkan, Astebro and Sierra (2016)

[12] The information was obtained from https://rpubs.com/dansc/kick on April 10, 2018. The data at RPubs is gathered by querying the (undocumented) Kickstarter API.

report an average campaign goal of GBP138,228 for projects at *Seedrs* (one of the largest UK equity CF sites). They also report a large heterogeneity and a steady increase over time (from GBP68,000 in 2012 to GBP200,000 in 2015).

Pledge levels. To provide some basic descriptive evidence about pledge levels in reward CF campaigns, we obtained data from Kickstarter. The data shows that the amount pledged per backer in reward CF campaigns is small, with a median of US$45.91 (mean US$66.49).[13]

Regarding equity CF, Vulkan, Astebro and Sierra (2016) report that the average individual pledge for successful campaigns run at the UK equity CF platform Seedrs is GBP368, compared to GBP233 for failed campaigns.

Therefore, *pledge levels in CF projects are very small, especially in reward CF,* which is an important piece of information for the debate on CF regulation.

Campaign length. Colombo, Franzoni and Lamastra (2015) report that the median duration of a Kickstarter campaign is 1 month, with about one half of all campaigns lasting 30 or 31 days, 25% being shorter, and the remaining campaigns lasting between 32 and 60 days.[14]

Campaign length has also been used as a control in regressions studying funding success, with an estimated coefficient generally negative (Mollick, 2014; Crosetto and Regner, 2014; Cumming, Leboeuf and Schwienbacher, 2015; Colombo, Franzoni and Lamastra, 2015).

Contracts. Wroldsen (2016) reviews the first wave of equity CF contracts in the US after the regulation of equity CF became effective in May, 2016 (see Section V for a discussion of the US regulation of CF). Wroldsen documents six general types of CF contracts: common stock (offered by 38% of companies), often issued as non-voting common, and generally without protections for investors; preferred stock (10%), whose characteristics vary greatly across offerings, but which are generally non-voting and provide different protections against future dilution of their investment; revenue-sharing agreements (8%), which are capped as a multiple of the initial investment in some cases; a type of convertible debt contract known as a KISS, for Keep It Simple Security (5%), which is essentially a debt contract that can be converted into different kinds of equity conditionally on the occurrence of different events (such as new rounds of financing or the achievement of valuation thresholds); SAFE (for Simple Agreement for Future Equity) contracts (31%), which, as their name suggest, are essentially deferred

[13] Data for the period 2009–2015. Accessed at https://rpubs.com/dansc/kick on April, 2018. See III.1.

[14] Kickstarter originally established a maximum campaign duration of 90 days, but low ered this maximum to 60 days in 2011.

equity investments whose terms depend on the valuation at the time the equity is issued; and interest-bearing loans (8%). Of the different kinds of contracts used, SAFEs and KISSes offer investors a greater array of protections than other contracts.

Therefore, although common stock is the most usual type of security in equity CF in the US, a significant variety of securities are used. Some of these securities are similar to securities used in angel or VC-funded projects, but generally award investor fewer protections and control rights.

In other countries, *contracts also differ in how the relation between in vestors and the issuer is structured by the platform.* Some platforms, such as UK platform Seedrs, set up a special purpose vehicle (SPV) that sits between investors and the issuer and then use "pooling contracts" to operationalize the investment. With this structure, the entrepreneur has to deal with only one shareholder of record. Other platforms, such as UK platform Crowdcube, offer backers direct participation in the securities. With either structure, contracts often include different kinds of provisions to protect investors. For example, Seedrs investors enjoy tag-along (the right to sell their participation at the same conditions as the majority shareholders, if they sell) and preemptive rights (the right to acquire any newly issued shares) that protect them against dilution. Crowdcube forces entrepreneurs to adopt one of two standardized templates of articles of incorporation, either creating a single class of stock for all investors or a tiered stock structure where more protected shares are only offered to large investors (Camara, 2016). Other contracts usually offered in Europe are SAFE contracts, silent partnerships (the equivalent of limited partnerships, where the investor does not participate in the operation of the firm, but shares in its profits), convertible bonds and, specially in Germany, profit-participating loans ("partiarisches Darlehen," where the lender receives a share in the firm's profit or revenue).

Interestingly, Wroldsen reports that in the US *many contracts also include non-monetary rewards.*

2. The Backers of Crowdfunding Projects

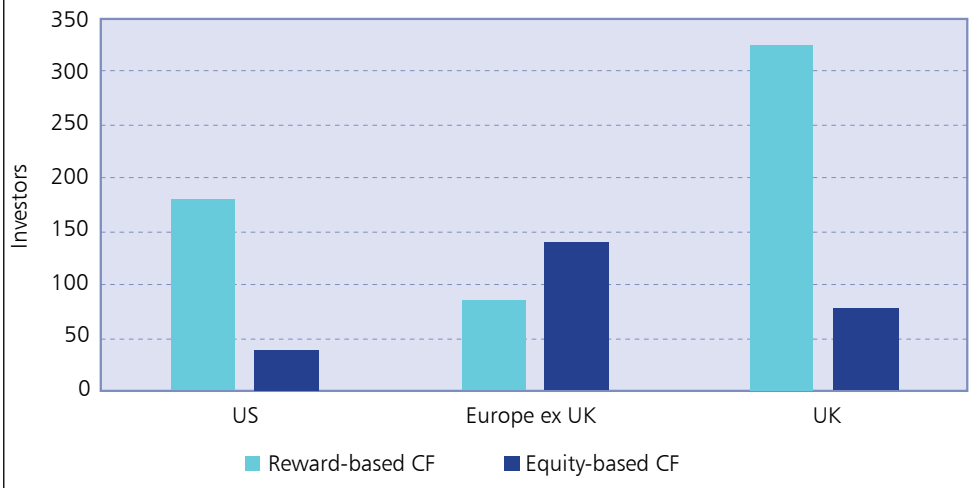
The identity and the screening and monitoring activities of backers are essential in determining the performance of a financing vehicle, as we discuss in Section II.3, as well as the need for regulation and the form of that regulation, as we discuss below in Section V. Although the theoretical literature on CF has paid little attention to these issues, the empirical work allows us to have a preliminary picture of who the backers are, what motivates them to participate in CF, and what they do.

How large is the crowd in CF? There is no authoritative information source that can answer this question globally, but there is information about individual countries and regions. For example, in the UK, alternative finance platforms reported that the number of users of the platforms grew by 13%, from 1.09 million in 2015 to over 2.5 million in 2016, although these figures surely suffer from high levels of double counting, as some investors will likely use more than one platform, and it is not restricted to CF investors exclusively (Zhang *et al.,* 2017). Wardrop *et al.* (2016) report a similarly calculated figure for the Americas: 9.7 million in 2015, 8.6 million of them from the US. It is interesting to compare theses numbers with existing estimates of the number of angel investors. For example, Huang *et al.* (2017) report that there were 297,880 angel investors in the US in 2016.

Looking at individual CF platforms also allows us to gauge the size of the backer population. Kickstarter, the leading reward-based CF platform, launched in April 2009 and by 2014 it had already received pledges from 5.7 million backers. That number grew to 14 million in April 2018.[15] Indiegogo, in turn, boasts 9 million backers on the same date.[16] In comparison, Crowdcube, one of

FIGURE 6

NUMBER OF INVESTORS PER DEAL



Note: Average number of investors per deal by CF type for selected regions for2015.

Sources: US (Ziegler *et al.,* 2017), Europe (Ziegler *et al.,* 2018) and the UK (Zhang *et al.,* 2017).

[15] Kickstarter's data accessed at https://www.kickstarter.com/help/stats on April, 2018.

[16] https://www.indiegogo.com/about/our-story accessed April 2018.

the leading equity CF platforms in the UK, increased its user base from around 10,000 in 2012 to more than 100,000 in 2015, and more than 400,000 by June 2017 (Estrin, Gozman and Khavul, 2018).

The above figures give an indication of the number of potential CF participants. However, it is also important to understand how large is the crowd that invests in any given project. Kickstarter data shows the median number of backers for all projects is 26 (mean 134.7).[17] Figure 6 displays the average number of investors per deal in 2015 for key regions per CF type. Reward CF attracts roughly four times as many investors per deal in the US and UK, but less investors per deal in Continental Europe (EU ex UK).

For equity CF, Estrin, Gozman and Khavul (2018) report that there are on average between 200 and 250 investors per successful campaign on UK equity CF sites.

The above data indicates that, indeed, *successful CF campaigns attract many more investors than any other traditional form of early stage financing.* At the same time, the data should also help calibrate the theoretical models to gauge the potential benefits stemming from the aggregation of dispersed information (how much information is likely to be obtained given the number of backers?) as well as backers' expectation of being pivotal.

One of the purported benefits of CF is the democratization of the investment in early stage companies, making it accessible to investors that were previously shut out of the market. However, a substantial proportion of equity CF investment comes from institutions or wealthy individuals, who would be able to invest through angel networks or VC funds (see Zhang *et al.,* 2017; Ziegler *et al.,* 2017, for the US). This proportion is especially large (71%) in the US because of regulatory restrictions, as we discuss in greater detail in Section V (Ziegler *et al.,* 2017).

Expectedly, the presence of professional investors in reward CF appears to be limited. Kuppuswamy and Bayus (2018) find that 72% of Kickstarter project supporters are one-time backers and likely from the sponsor's social circle (95% of total one-time backers).[18]

Where are the backers? Angel investors and VCs tend to invest in geographically close ventures. It is important to understand whether CF investors

[17] Data for the period 2009–2015. Accessed at https://rpubs.com/dansc/kick on April, 2018. See III.1.

[18] Kickstarter reports 4.7 million repeat backers out of a total of 14.5 million (https://www.kickstarter.com/help/stats accessed on April 14, 2018).

also invest locally. Agrawal, Catalini, and Goldfarb (2015) study SellaBand (an equity CF platform that connected musicians with investors) and find that the average distance between an artist and an investor was approximately 5,000 km. They report that local and far-away investors are qualitatively different in that local investors are less responsive to information on cumulative funds raised. However, this difference is explained by pre-existing social relationships with the entrepreneur ("friends and family"), which are disproportionately local in nature. When comparing the concentration of reward-based CF projects to that of VC, Mollick (2013) finds that both have a similar degree of clustering, but CF is slightly less concentrated than VC, and CF and VC projects are not clustered in the same areas. Therefore, *existing evidence suggests that the geographic reach of CF is wider than that of alternative sources of early stage financing. However, local investors still play an important role.*

What are the backers' motivations? *The consumption of the product developed by the entrepreneur is the most important motivation for the backers of reward CF projects. However, a subgroup of backers also report to value other dimensions of CF,* such as the involvement in the project, being part of a community, engaging in innovative behavior, the ability to contribute to a larger goal or do good, and the possibility of supporting a particular person or group (Steigenberger, 2017; Gerber and Hui, 2013; Ordanini *et al.,* 2011; Kuppuswamy and Bayus, 2017). Steigenberger (2017) finds that those that report supporting a particular sponsor or goal as a significant motivation contribute significantly more. Zheng *et al.* (2014), Colombo *et al.* (2015), and Skirnevskiy, Bendig, and Brettel (2017) also show that reciprocity in contributions matters. For example, Colombo *et al.* (2015) find that the number of Kickstarter projects a sponsor had backed before launching her own campaign significantly increases the number of early backers and the amount contributed by early backers.

In equity CF, financial returns are the main motivation of backers. For example, Zhang *et al.* (2017) find that of 88% of the UK equity CF investors surveyed regarded making a financial return as important to very important, and 81% view the ease of the investment process as key in their decisions. *However, more than 50% of respondents report non-financial factors as important or very important drivers of their decisions* (such as investing in industries they know or care about, knowing that their money is helping a business, supporting an alternative to big banks, feeling their money is making a difference, or curiosity). A significant 24% of respondents find it important or very important to support a friend or family member.

What is the screening and due diligence process like? Angel investors and VCs listen to entrepreneurs' pitches, interview them, read their business plans, and perform other due diligence tasks. What do CF backers and platforms do?

Survey evidence suggests that *equity CF investors conduct little formal screening of their investments.* Zhang *et al.* (2017) report that 42% of polled UK equity CF investors said they spent at most 20 minutes per week picking potential investments. 57% of respondents relied on the equity CF platform for due diligence, and only 26% performed the analysis themselves. Most investors (83%) expect platforms to verify basic information about the company seeking finance, its financial information, and how it intends to use the funds. Indeed, regulation requires equity CF portals to conduct significant due diligence and they report that they perform such due diligence.[19] Crowdcube, one of the leading UK platforms, states that its due diligence process can take between 3 to 4 weeks, but can be longer. Reward CF platforms, however, conduct only minimal due diligence and often rely on the information provided by platform users to detect possible cases of fraud or sponsors that do not abide by the platform's rules.

One can obtain some indirect evidence about the screening carried out by CF investors by analyzing what project characteristics that investors can observe at the time of the investment decision are associated with funding success.

Equity CF investors value projects in which the stake retained by the entrepreneur is large, projects with larger boards, and projects with MBA graduates as executive board members, but do not seem to be influenced by project size, the number of planned years to exit, and external certification stemming from patents, grants or awards (Ahlers *et al.,* 2015; Vismara, 2016).

Backers in reward CF seem to pay attention to and value past performance in CF campaigns (Courtney, Dutta, and Li, 2017; Skirnevskiy, Bendig, and Brettel, 2017; Buttice, Colombo, and Wright, 2017). Mollick (2014) finds evidence that signals such as videos and frequent updates are significantly associated with greater success, and spelling errors significantly reduce the chance of success, which suggests that *backers pay attention to the materials offered by sponsors when deciding whether to contribute.* Courtney, Dutta, and Li (2017) also reports that media usage helps explain campaign success.

Backers also seem to base their decisions on other backers' decisions and comments. In the subsection on campaign dynamics below, we discuss how the path of contributions during a campaign determines subsequent contributions. Here we note that several papers analyzing reward CF also document that

[19] For a description of the due diligence practices of Seedrs and Crowdcube, two of the main equity CF platforms in the UK, see https://www.seedrs.com/learn/wp- content/uploads/2017/08/Seedrs-Standard-Guide-to-Due-Diligence.pdf and https://help.crowdcube.com/hc/en-us/articles/206234044-What-is-Crowdcube-s-equity-crowdfunding-due-diligence-process-, respectively (accessed in April, 2018).

comments made by other backers, both during the campaign (Stanko and Henard, 2017; Bi, Liu, and Usman, 2017) and during previous campaigns (Li and Martin, 2016), matter for campaign success.

As discussed above, Agrawal, Catalini, and Goldfarb (2015) argue that local investors rely on local information, because they tend to be less sensitive to the amount of funding provided by other backers. Relatedly, several authors find that *different measures of online "social capital"* (such as the number of Facebook friends or LinkedIn links) *have an impact on project success both for reward CF* (Mollick, 2014, and Li and Martin, 2016, although in the latter case, the estimated effect is small) *and for equity CF* (Vismara, 2016). The influence of the size of the founder's online network on project success may indicate that some pre-existing connection to the founder helps screen projects, in line with findings in the entrepreneurial finance literature that show that social contacts help overcome the information asymmetries between entrepreneurs and investors (Shane and Cable, 2002). However, a greater network may be indicative of underlying founder characteristics that are conducive to success or of a larger number of people who may contribute to the project with the goal of supporting the founder. Zheng *et al.* (2014) compare the effect of social capital in US and China and find that it is stronger in China.

How effective are CF backers in identifying value creating projects? Mollick and Nanda (2015) use an experimental design that compares funding decisions for proposed theater projects by a panel of US experts and data from Kickstarter and find that in 59% of cases both agree. Furthermore, they argue that their data suggests that CF has the potential to reduce "false negatives" (that is, that it may fund viable projects that are rejected by experts), since they find that for 75% of the projects for which there was disagreement between experts and the crowd, the crowd financed projects that experts rejected and that *ex post* could be considered as successful as accepted projects.[20]

3. Entrepreneurs and Projects

Why do entrepreneurs use CF to finance their ventures? Mollick and Kuppuswamy (2014) conducted a survey of sponsors of Kickstarter projects and report that almost 70% of successful sponsors (60% of unsuccessful ones) wanted to see if there was demand for their product, 65% saw CF as a way of marketing their product, and 20% wanted to get ideas on how to improve their product. *The survey evidence thus confirms the attention paid by the theoretical*

[20] In particular, these projects were equally likely to be on budget, had commercial success, and received positive critical reviews.

literature to the role of reward CF as a tool to learn about demand and market a product. Mollick and Kuppuswamy (2014) also report that over 50% of sponsors wanted to connect directly with a community and 59% saw their project as a first step towards launching a company. Similar findings are reported by Stanko and Henard (2017) and Gerber and Hui (2013), although the latter also list the ability to get funds when traditional sources are dry as one of the main motives for trying to raise funds through CF.

Spreading awareness of their products and ease of application were also identified by Estrin Gozman, and Khavul (2018) in their qualitative study as motives for entrepreneurs to seek equity CF as an alternative to traditional funding. Other motivations for equity CF sponsors were obtaining financing without relinquishing as much control as with VC and the digitization (and consequent simplification) of the pitching process. Estrin Gozman, and Khavul (2018) also report that entrepreneurs mentioned the risk of publicly failing to secure financing and the problem of having to deal with unsophisticated investors as factors keeping them away from equity CF.

It is worth noting that *some entrepreneurs use CF as a sustainable way to access finance.* Thus in Europe (ex. UK) in 2016, 11% of reward-based CF sponsors and 10% of equity CF sponsors were repeat sponsors (Ziegler *et al.,* 2018). Relatedly, it is important to understand whether CF is used to fund one-time projects only or as part of the development of entrepreneurial firms. In a survey of successful Kickstarter projects, Mollick (2016) find that the percentage of campaigns that raise money for firms or institutions with the intention of continuing operations is relatively small for art-oriented categories but very significant (between 45% and 65%) for product-oriented categories. *Thus, at least in some product categories, reward CF sponsors raise funds with the intention of financing the development of firms and not just one-time projects.*

We know relatively little about sponsors and their paths to entrepreneurship. Mollick (2016) report that *most sponsors are young (25-34 years old) and highly educated (82% have a college degree, of which 34% also have an advanced degree). Sponsors are not wealthy at the time they initiate their campaigns.* Thus Mollick (2016) report that sponsors' mean earnings before the campaign were over US$48,000 but 16% of all sponsors reported earnings of less than US$10,000. 39% of creators were employed full–time, and 19% were freelancers.

At what stage and for which kinds of innovations do entrepreneurs seek CF? The existing evidence suggests that *potential equity CF backers prefer to fund more advanced projects than reward CF backers* (Stanko and Henard,

2017). At the same time, *reward CF backers prefer incremental innovation projects (which are less risky, easier to understand, and more susceptible of improvement through backers' inputs) over more radically innovative ones* (Chan and Parhankangas, 2017; Gerber and Hui, 2013).

4. Campaign Dynamics

CF campaigns take place over a period of time during which potential backers can observe the path of contributions to the campaign. As we discuss in our review of the theoretical papers on herding, the ability of backers to condition their contribution to a campaign on the history of contributions to that campaign may lead to the emergence of herding behavior and, even, informational cascades. Hornuf and Schwienbacher (2015), Vismara (2015), and Astebro *et al.* (2018) report *behavior consistent with informational herding in equity CF campaigns, that is, with backers sometimes disregarding their private information about project value and deciding instead on the basis of the path of the amount contributed by other backers.*

Vulkan, Åstebro and Sierra (2016) find that a *few large investments have a major role in funding success.* Thus, the largest pledge accounts for 30% of the funding goal for successful projects, and for 5.4% for failed campaigns, a result which they interpret as evidence that the fact that some backers commit substantial resources may serve as a signal of the quality of the project for other backers.

Using Kickstarter data, Kuppuswamy and Bayus (2018) find that *the typical pattern of backer contribution is U-shaped, with backers more likely to contribute towards the beginning and the end of the campaign.* Project updates have a positive effect on backer contribution and they themselves follow a U-shape. Skirnevskiy, Bendig and Brettel (2017) find evidence that *there is a higher relative share of funding from loyal backers in the early campaign period than in later periods,* and Agrawal, Catalini and Goldfarb (2015) reports that *the social network of sponsors is the initial source of significant funding for many projects, the equivalent of "friends and family".*

Kuppuswamy and Bayus (2018) and Kuppuswamy and Bayus (2017) also find evidence of a "goal gradient" effect: backers support CF projects when they believe that their contribution will make an impact. Thus, *support for a campaign increases as it approaches its funding target and decreases after the target is reached.* The goal gradient effect is accentuated as the project deadline approaches and for projects that do not obtain a lot of early support.

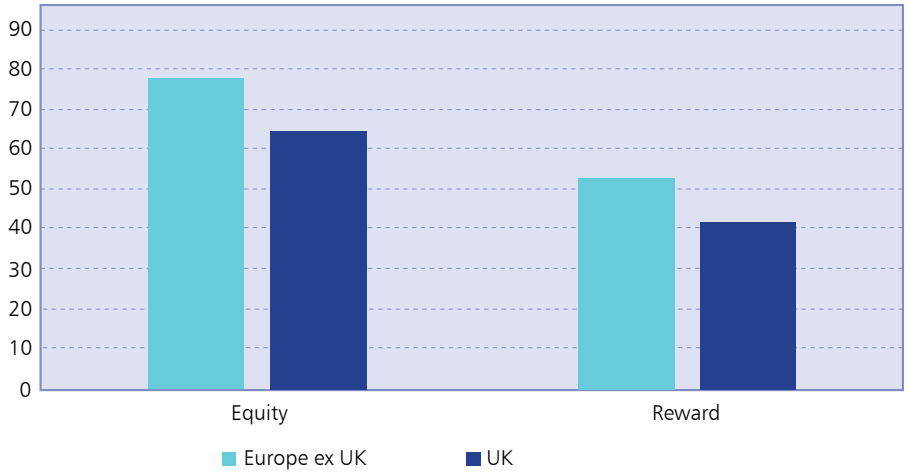
5. Crowdfunding Campaign Outcomes

Funding success. Our discussion above already mentions several factors that contribute to project funding. Here we focus on the distribution of project funding, since it can shed light on the motivations of backers and can be a testing ground for theories of CF.

Figure 7 *shows that funding success rates vary significantly by CF type and jurisdiction*.

FIGURE 7

COMPARISION OF SUCCESS RATES (%)



Note: Funding success rates for 2016 by type of crowdfunding for selected regions.

Sources: Europe (Ziegler *et al.,* 2018), UK (Zhang *et al.,* 2017).

Crowdfunding campaigns that fail do so by large amounts, while successful ones are so by small amounts. This has been shown in both equity (Vulkan, Astebro and Sierra, 2016) and reward CF (Mollick, 2014). Thus, amongst Kickstarter failed projects, mean amount funded is 10.3% of the goal. In contrast, twenty five percent of projects that are funded are 3% or less over their goal, and only 50% are about 10% over their goal. However, Chemla and Tinn (2017) report that there is significant overfunding for Kickstarter projects in the technology, product design, or software categories.

Longer term success. Mollick and Kuppuswamy (2014), in a survey of successful Kickstarter projects, find that over 90% of successful Kickstarter

projects remained ongoing enterprises, with 32% of them reporting yearly revenues of over US$100,000, a year after the campaign. Mollick (2016) estimate that one dollar of pledges generates an average of US$2.46 in revenue outside of Kickstarter. *The little evidence available so far thus suggests that successful reward CF projects have a high survival rate.*

Delivery of rewards. As we discuss in our review of the theoretical literature, the fact that the sponsors of reward CF projects do not have clear contractual obligations to deliver rewards creates opportunities for reneging on the promises to deliver. However, Mollick (2014) reports that *fraud amongst Kickstarter campaigns is very rare* (the direct failure rate was only 3.6%), despite the platform having no effective mechanism to enforce the delivery of rewards. *A majority of the projects however, suffer delays* (of 2.4 months on average), with slightly below 25% of projects delivering their rewards on time, and with larger and more overfunded projects suffering longer delays. Despite these pervasive delays, over 75% of backers say they where happy with the outcome and over 35% found results to be better than expected (Mollick and Kuppuswamy, 2014).

6. Challenges and Suggestions for Future Empirical Research

Since the first theoretical papers have been published only recently, the empirical work on CF has evolved mostly without reference to formal theory. Future work testing the predictions of existing theoretical models could contribute greatly to our understanding of CF. More generally, the interpretability of the empirical findings would be greatly enhanced if the empirical methodology were informed by theory to a larger extent.

The generalizability of the empirical findings about CF is a major concern for at least two reasons. First, CF is still a new phenomenon, so it is not clear whether the results obtained for the first cohorts of projects and investors will survive as the sector matures. Second, most empirical studies rely on data from one or two platforms and, often, only for a subsample of projects. We believe that much could be gained by analyzing data from different platforms side by side and by analyzing longer sample periods.

Another area that has not received the necessary attention yet has been the measurement of the risk and return of funded CF projects. Although there is some preliminary evidence on the exit paths and financial performance of crowdfunded firms, this question should receive much greater attention as data on crowdfunded firms accumulates.

There are many open questions having to do with the relation between CF and other sources of entrepreneurial financing. Understanding how the projects financed by CF differ from those financed by angel investing and VC would help address these questions.

Another key and under-researched area is the impact of CF on the transition to entrepreneurship. To what extent is CF allowing would-be entrepreneurs to start new companies? Where are these entrepreneurs coming from? How may CF alter human resource management practices in small and large innovative companies?

As we discuss in Section V, one of the main concerns of regulators is the possibility that small investors may underestimate the risks of CF. To address this concern, more research is needed to ascertain whether CF backers are investing optimally. On a related note, it is also important to understand whether the money invested in crowdfunded projects is flowing out of other forms of financing innovation, other types of assets, or it is new net saving.

Finally, a whole set of issues concerning the non-monetary returns to CF is left unanswered. If backers value the social return of a crowdfunded project, what is the price they attach to it? Is their "investment" in prosocial CF a substitute for other forms of prosocial investment or charitable giving?

V. THE REGULATION OF CROWDFUNDING

Despite their commonalities, equity CF and reward CF fall within very different regulatory realms. In equity CF, investors are promised some kind of financial return by the entrepreneur. Therefore, the contract between investors and the entrepreneur is considered a security (US) or a financial instrument (EU), so that equity CF falls within the realm of securities regulation.[21] In contrast, since reward CF does not promise a monetary return to backers, but, rather, some non-monetary reward (often the good produced by the entrepreneur), it is generally not considered to be covered by securities law but by general contract law and by consumer protection regulation (Armour and Enriques, 2018). Another stark difference between the regulation of equity and reward CF has been the extent of regulatory action directed at each type of CF: whereas equity CF has been explicitly regulated in numerous jurisdictions, there is little regulation directed at reward CF.

[21] See Bradford (2012) for a detailed discussion of the regulation of CF in the context of securities law.

1. The Regulation of Equity Crowdfunding

1.1. Securities Regulation

The goal of securities regulation is to ameliorate the asymmetric information and moral hazard problems that characterize the issuance and trading of securities. Securities regulation tries to achieve this goal through two main kinds of instruments: information disclosure requirements and the regulation of the conduct of the financial intermediaries that assist with the issuance and trading of securities.

Information disclosure requirements for the issuer take the form of the obligation to register the security with the regulator, an obligation that entails the provision of detailed information about the company issuing the security and its management. The costs of the registration process (which include the fees paid to regulators and accountants, as well as underwriter fees), however, are substantial, and the process takes time. Beyond the costs at the time of issuance, registration also entails ongoing information disclosure requirements (among them, the need to provide full financial statements reviewed by certified accountants periodically). As a result, security registration is typically not feasible for firms seeking seed or early stage financing.

Entrepreneurial firms seeking to obtain funding from several investors can do so by issuing securities that benefit from some exemption from the registration requirements. These exemptions mainly involve strict restrictions on the publicity of the offering (*i.e.,* requiring the offering to be "private"), the requirement that investors be "accredited" or "qualified," and limits to the size of the offering. Thanks to these exemptions, early-stage firms have been able to obtain financing from qualified angel investors and venture capital funds. It is important to note that, generally, individual investors are considered to be accredited if their income or wealth are sufficiently high (although in some jurisdictions, investors may qualify if they have enough prior experience with private equity).[22] The justification of this definition of accredited investor is that wealthy individuals have the means to obtain the necessary investment advice and the ability to diversify their investments.

[22] For example, in the US an individual is considered to be an accredited investor if his or her net worth (or joint net worth with a spouse) exceeds $1 million or if his or her annual income exceeded $200,000 in each of the two years prior to the investment (or the joint annual income with a spouse exceeded $300,000 for those years), with a reasonable expectation of the same income level in the year of the investment. In the UK, having over GBP250,000 in assets (excluding the value of one's primary residence and pension assets) or income over GBP100,000 in the last year would qualify an investor for participation in unregistered offerings. In Spain, accredited investors must have an annual income in excess of €50,000 or a net worth in excess of €100,000.

The other aspect of securities regulation that is relevant for CF concerns the requirements imposed on brokers or investment advisers, which take the form of disclosure requirements (through mandatory registration) as well as requirements concerning their behavior as agents of investors. For example, brokers owe their clients different duties having to do with the accurate representation of information or with the suitability of investments for their clients. Brokers must also charge reasonable prices and ensure the best execution of the clients' orders. In some jurisdictions, such as the US, investment advisers have strong fiduciary duties towards investors. Since the activities of equity CF portals fall within those of brokers and/or investment advisers, in the absence of specific exemptions they would be subject to the regulation covering these financial intermediaries.[23]

1.2. The Regulation of Equity Crowdfunding

In some jurisdictions (notably, the UK and Germany), the registration exemptions available to entrepreneurs prior to the arrival of CF provided the necessary room for equity CF to develop. In others (notably, the US), however, the conduct of equity CF was largely inconsistent with existing exemptions, so that new specific exemptions have been introduced to avoid registration of crowdfunded securities.[24] Still in others countries (notably China), no new exemptions have been introduced, so online equity offerings are restricted to accredited investors (Garvey *et al.,* 2017).

The new exemptions take mainly two forms. The first one (adopted, for example, by Title II of the US JOBS Act), lifts most restrictions on the publicity of the offering as long as investors are accredited. This form of financing may be considered "crowdfunding", since the use the web to solicit investment may allow a large pool investors to participate in the offering, even if the accreditation requirement reduces the size of the potential "crowd." Because of this requirement, however, this form of financing is perhaps better understood as lying between angel investing and CF.

The second kind of new exemption from registration allows for the participation of unaccredited investors but (a) imposes additional disclosure, and even –relatively mild– registration requirements on issuers, and (b) limits the amount that can be raised (for example, to be below US$1 million over a

[23] See Bradford (2012) for a discussion of the categorization of CF portals as brokers or investment advisers in the US context, and European Commission (2016) in the context of the European Union.

[24] In the US, these exemptions were introduced by the Jumpstart Our Business Startups (JOBS) Act of 2012. New regulations were also introduced in several EU countries, such as Italy (in 2012), France (in 2014), or Spain (in 2015).

year, in the US case) and, importantly, the amount that unaccredited investors are allowed to invest.[25]

One of the main issues in the regulatory debate is the effectiveness of information disclosure given small investors incentives to read and process the information provided by issuers, and with their ability to understand the risks of CF investments. Empirical work, especially lab or field experiments, is needed to shed light on these questions.

In order to make the restrictions on the participation of unaccredited investors meaningful, new exemptions often restrict temporarily the resale of crowdfunded securities to such investors. More research comparing the potential benefits of the constraints on unaccredited investors with the potential costs of not having a secondary market for crowdfunded securities would be welcome to inform the regulatory debate.

1.3. The Regulation of Platforms

Regulation also imposes disclosure and conduct requirements on CF platforms. For example, US regulation requires equity CF offerings to unaccredited investors to be offered through a registered broker-dealer or a registered *funding portal,* which can be considered a limited-purpose form of broker, which, because of the limitations on its activities, faces milder registration requirements and other constraints than brokers.

In the European Union (EU), equity CF platforms generally have to be authorized under the Markets in Financial Instruments Directive (MiFID), although there is some variation between member states in the precise implementation of MiFID rules. EU regulation also imposes minimum capital requirements on platforms.

There is a wide variety of other requirements imposed on CF intermediaries, which vary by jurisdiction. For example, the intermediary may be required to

[25] For example, Title III of the US JOBS Act, (known as the "Capital Raising Online While Deterring Fraud and Unethical Non-Disclosure" Act or CROWDFUND Act), which became effective in 2016, establishes that an investor whose annual income and net worth are both below US$100,000 cannot invest in a single issuer more than US$2,000 or 5% of his annual income or net worth and sets less strict limits for investors who have either net worth or annual income above US$100,000. In the EU, most countries have some kind of limitation on investable amounts, ranging from 1,000 euros in Belgium, to 3,000 euros per project or 10,000 per year in Spain. In Germany the limit per issuer is set to twice the monthly income or 1,000 euros if the investor is not willing to disclose the necessary information. See Hornuf and Schwienbacher (2017) for a review of the regulation.

have a reasonable basis to believe that issuers and investors satisfy regulatory requirements (having to do with their income, wealth, or investment limits), ensure the suitability of the investment for investors, provide educational materials to investors, provide public communication channels for investors and representatives of the issuer to interact and exchange comments. Other noteworthy requirements are the requirements to conduct AoN campaigns (and not KIA campaigns) or the prohibition to have a financial interest in issuers using their platforms. The latter prohibition may limit the possible role of CF intermediaries in the screening and monitoring of crowdfunded projects and deserves more theoretical attention. Interestingly, there is relatively little regulation on the ways in which platforms may be compensated by sponsors or investors (although in some cases, platforms are not allowed to receive securities as compensation). Given that different kinds of compensation may provide platforms different incentives for screening and for the design of CF campaigns, we believe that it is worth investigating, both theoretically and empirically, the impact that different compensation schemes may have on platform behavior and whether there may be a need to regulate platform compensation.

2. The Regulation of Reward Crowdfunding

In reward CF, backers are not promised any financial return. Therefore, reward CF does not fall within the reach of securities regulation. Armour and Enriques (2018) argue that, as long as the product or service developed by the entrepreneur is offered as a reward, general contract law and consumer protection obligations cover reward CF. These authors argue as well that the strength of these obligations is very different across jurisdictions. In particular, US consumer regulation is relatively lax and allows parties to waive some protections by contract. Thus, for example, sponsors may not be liable for the late delivery of a product offered as a reward, as long as the sponsor exerted "best effort" and clearly communicated with backers. In contrast, Armour and Enriques (2018) argue that if the product is offered as a reward, in the EU stricter regulations cover reward CF, which give consumers the non-waivable right to cancel a purchase (and receive a refund) within fourteen days of its reception. Further, sponsors are subject to stronger liability if they omit material information or offer misleading information, and need to satisfy standards of fairness in their contracts with backers, which Armour and Enriques (2018) argue may be violated by non delivery. These authors argue that EU regulation fails to accommodate the fact that reward CF is not simply a pre-sale, but, rather, a financing contract in which the backer expects to share the risk that the product cannot be developed in time, with the expected characteristics, or at all. As discussed in section III.3, Gutiérrez Urtiaga and Lacave Sáez (2018) argue, in a different vein, that strict penalties for non-delivery may limit the ability of reward CF to serve as a credible signal of entrepreneurial ability.

VI. CONCLUDING REMARKS

Despite being a very recent phenomenon, there is already a sizeable, and rapidly growing, literature on CF. In this review, we have tried to provide an organized account of what we know about CF and of what we do not know but would be important to know.

The theoretical analysis of CF in economics and finance has, understandably, focused on the ability of CF to "harness the wisdom of the crowd", that is, to obtain information dispersed among would-be consumers or investors to guide entrepreneurial decisions. Initial theoretical results about reward CF generally support this potential role of CF, whereas the results regarding equity CF are less clear. Some papers have also highlighted that CF may be a means to price discriminate prior to the production of the good. Moreover, such price discrimination may be efficient in some cases, since it may allow for the efficient provision of the good in contexts in which traditional financing would not have been possible.

Our review of the theoretical literature shows that there are key questions that have received little attention and that would benefit from a theoretical analysis. Among them, we highlight the analysis of CF contract design, a more careful modeling of the incentives and behavior of CF platforms, a greater integration with the research on angel investing and venture capital, and the explicit inclusion of the kinds of limitations of small investors that motivate regulators' concerns about CF.

The empirical analysis carried out so far allows us to have a reasonably clear picture of the phenomenon of CF, but there is much work to be done. Perhaps the most pressing task is to gather rich cross-platform data sets covering relatively long periods. We think that future empirical work should be more firmly grounded in theory to be able to address some of the main unresolved questions regarding CF, such as whether it may be able to fill a "funding gap" for at least some kinds of small, risky projects, not financed by angel investors or VCs, or whether and how it should be regulated. Although the empirical work has had a wider reach than the theoretical analysis, the areas that we identify above as needing more theoretical attention are also areas in which the returns of empirical work would be large. More work on the identity of project backers and the role played by different types of backers would also be useful. Because of the size and particular features of the Chinese economy and CF sector, more empirical work should focus on Chinese CF. Analyzing the potential and challenges of CF in economies with less developed financial sectors appears to us as another promising avenue for future research.

To conclude, we would like to emphasize that whereas some forms of CF are consolidating in some regions, for example reward CF in the US or equity CF in the UK, in other regions they are very much in flux. This is especially the case for equity CF, whose legal status has been clarified only very recently in a number of countries, and which is still highly uncertain in key economies such as China. At the same time, new forms of CF, such as Initial Coin Offerings (ICOs), have experienced explosive growth. The rapid transformation of crowdfunding and, more generally, internet-enabled financing is sure to raise new and exciting questions for researchers.

BIBLIOGRAPHY

Agrawal, A.; Catalini, C., and A. Goldfarb (2014), "Some simple economics of crowdfunding," *NBER Innovation Policy & the Economy (University of Chicago Press),* 14 (1): 63–97.

— (2015), Crowdfunding: "Geography, social networks, and the timing of investment decision," *Journal of Economics & Management Strategy,* 24 (2): 253–274.

Ahlers, G. K. C.; Cumming, D.; Guenther, C., and D. Schweizer (2015), "Signaling in equity crowdfunding," *Entrepreneurship Theory and Practice,* 39: 955–980.

Armour, J., and L. Enriques (2018), "The promise and perils of crowdfunding: Between corporate finance and consumer contracts," *The Modern Law Review,* 81 (1): 51–84.

Armstrong, M. (2006), "Competition in two-sided markets," *The RAND Journal of Economics,* 37 (3): 668–691.

Astebro, T. B.; Fernández Sierra, M.; Lovo, S., and N. Vulkan (2018), Herding in equity crowdfunding, *SSRN Scholarly Paper ID* 3084140, Social Science Research Network, Rochester, NY.

Babich, V.; Tsoukalas, G., and S. Marinesi (2017), Does crowdfunding benefit entrepreneurs and venture capital investors?, *SSRN Scholarly Paper ID* 2971685, Social Science Research Network, Rochester, NY.

Bachman, A.; Becker, A.; Buerckner, D.; Hilker, M.; Lehmann, F.; Tiburtius, P., and F. Burkhardt (2011), "Online peer-to-peer lending: A literature review," *Journal of Internet Banking and Commerce,* 162 (2): 1–18.

Banerjee, A. V. (1992), "A simple model of herd behavior," *The Quarterly Journal of Economics,* 107 (3): 797–817.

BBC News (2013), The Statue of Liberty and America's crowdfunding pioneer, *BBC News.*

Belleflamme, P.; Lambert, T., and A. Schwienbacher (2014), "Crowdfunding: Tapping the right crowd," *Journal of Business Venturing,* 29 (5): 585–609.

Belleflamme, P.; Omrani, N., and M. Peitz (2015), "The economics of crowdfunding platforms," *Information Economics and Policy,* 33: 11–28.

Bi, S.; Liu, Z., and K. Usman (2017), "The influence of online information on investing decisions of reward-based crowdfunding," *Journal of Business Research,* 71: 10–18.

Bikhchandani, S.; Hirshleifer, D., and I. Welch (1998), "Learning from the behavior of others: Conformity, fads, and informational cascades," *Journal of Economic Perspectives,* 12 (3): 151–170.

Bradford, C. S. (2012), "Crowdfunding and the federal securities laws," *Columbia Business Law Review,* 2012 (1).

Bradley, D. J.; Gonas, G. S.; Highfield, M. J., and K. D. Roskelley (2009), "An examination of IPO secondary market returns," *Journal of Corporate Finance,* 15 (3): 316–330.

British Business Bank (2017), Small Business Equity Tracker, *Technical report,* British Business Bank.

Brown, D. C., and S. Davies (2018), Financing efficiency of securities-based crowdfunding, *SSRN Scholarly Paper ID* 2692828, Social Science Research Network, Rochester, NY.

Burniske, C., and J. Tatar (2018), *Cryptoassets: The Innovative Investor's Guide to Bitcoin and Beyond,* New York, McGraw Hill Education.

Buttice, V.; Colombo, M. G., and M. Wright (2017), "Serial crowdfunding, social capital, and project success," *Entrepreneurship Theory and Practice,* 41: 183–207.

Camara, A. (2016), "Anonymous capital: Managing shareholder volume for equity crowdfunded companies in Canada," *Banking & Finance Law Review,* 31 (2): 259.

Casamatta, C. (2003), "Financing and advising: Optimal financial contracts with venture capitalists," *The Journal of Finance,* 58 (5): 2059–2085.

Chakraborty, S., and R. Swinney (2017), Signaling to the crowd: Private quality information and rewards-based crowdfunding, *SSRN Scholarly Paper ID* 2885457, Social Science Research Network, Rochester, NY.

Chan, C. S. R., and A. Parhankangas (2017), "Crowdfunding innovative ideas: How incremental and radical innovativeness influence funding outcomes," *Entrepreneurship Theory and Practice,* 41: 237–263.

Chang, J.-W. (2016), The economics of crowdfunding, *SSRN Scholarly Paper ID* 2827354, Social Science Research Network, Rochester, NY.

Chemla, G., and K. Tinn (2017), Learning through crowdfunding, *SSRN Scholarly Paper ID* 2796435, Social Science Research Network, Rochester, NY.

Chen, R.; Gal-Or, E., and P. Roma (2017), Reward-based crowdfunding campaigns: informational value and access to venture capital, *Working Paper,* available at https://www.idc.ac.il/he/schools/economics/research/Documents/crowdmodelname.pdf.

Cole, R. A., and T. Sokolyk (2017), "Debt financing, survival, and growth of start-up firms," *Journal of Corporate Finance,* in press.

Colombo, M. G.; Franzoni, C., and C. Rossi Lamastra (2015), "Internal social capital and the attraction of early contributions in crowdfunding," *Entrepreneurship Theory and Practice,* 39 (1): 75–100.

Cong, L. W., and Y. Xiao (2018), Up-cascaded wisdom of the crowd, *SSRN Scholarly Paper ID* 3030573, Social Science Research Network, Rochester, NY.

Cornelli, F. (1996), "Optimal selling procedures with fixed costs," *Journal of Economic Theory,* 71 (1): 1–30.

Courtney, C.; Dutta, S., and Y. Li (2017), "Resolving information asymmetry: Signaling, endorsement, and crowdfunding success," *Entrepreneurship Theory and Practice,* 41: 265–290.

Crosetto, P., and T. Regner (2014), "Crowdfunding: Determinants of success and funding dynamics," *Jena Economic Research Papers,* 035.

Cumming, D. J.; Leboeuf, G., and A. Schwienbacher (2015), Crowdfunding models: Keep-it-All vs. All-or-Nothing, *SSRN Scholarly Paper ID* 2447567, Social Science Research Network, Rochester, NY.

DeGennaro, R. P. (2012), Angel investors and their investments, in Oxford *Hand book of Entrepreneurial Finance:* 392–423, Oxford University Press.

EBAN (2016), European early stage market statistics, *Technical report,* EBAN The European Trade Association for Business Angels, Seed Funds and Early Stage Market Players.

Ellman, M., and S. Hurkens (2016), Optimal crowdfunding design, *SSRN Scholarly Paper ID* 2733537, Social Science Research Network, Rochester, NY.

Estrin, S.; Gozman, D., and S. Khavul (2018), "The evolution and adoption of equity crowdfunding: Entrepreneur and investor entry into a new market," *Small Business Economics:* 1–15.

European Commission (2016), Crowdfunding in the EU capital markets union, *Commission Staff Working Document SWD*(2016) 154 final, European Commission.

Garvey, K.; Chen, H.-Y.; B. Zhang; Buckingham, E.; Ralston, D.; Katiforis, Y.; Ying, K.; Deer, L.; Maddock, R., and T. Ziegler (2017), *Cultivating growth. The Asia Pacific Region Alternative Finance Industry Report 2*, Cambridge Center for Alternative Finance.

Gerber, E. M., and J. Hui (2013), "Crowdfunding: Motivations and deterrents for participation," *ACM Transactions on Computer-Human Interaction (TOCHI),* 20 (6): 34.

Goldfarb, B. D.; Hoberg, G.; Kirsch, D., and A. J. Triantis (2013), Are angels different? An analysis of early venture financing, *SSRN Scholarly Paper,* Social Science Research Network, Rochester, NY.

Gruener, H. P., and C. Siemroth (2017), Crowdfunding, efficiency, and inequality, *SSRN Scholarly Paper ID* 2886401, Social Science Research Network, Rochester, NY.

Gutiérrez Urtiaga, M., and I. M. Lacave (2018), The promise of reward crowdfunding, *SSRN Scholarly Paper ID* 3096753, Social Science Research Network, Rochester, NY.

Hakenes, H., and F. Schlegel (2014), Exploiting the financial wisdom of the crowd – crowdfunding as a tool to aggregate vague information, *SSRN Scholarly Paper ID* 2475025, Social Science Research Network, Rochester, NY.

Hellmann, T. (2006), "IPOs, acquisitions, and the use of convertible securities in venture capital," *Journal of Financial Economics,* 81 (3): 649–679.

Hornuf, L., and M. Neuenkirch (2017), "Pricing shares in equity crowdfunding," *Small Business Economics,* 48: 795–811.

Hornuf, L., and A. Schwienbacher (2015), Funding dynamics in crowdinvesting. Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische En twicklung – Theorie und Politik – Session: Financial Economics II B10-V3.

— (2017), "Should securities regulation promote equity crowdfunding?," *Small Business Economics,* 49: 579–593.

Huang, L.; Wu, A.; Lee, M. J.; Bao, J.; Hudson, M., and E. Bolle (2017), The American angel, *Technical report,* Angel Capital Association.

Ibrahim, D. M. (2008), "The (not so) puzzling behavior of angel investors," *Vanderbilt Law Review,* 61: 1405–1454.

Invest Europe (2017), 2016 European Private Equity activity, *Technical report,* Invest Europe.

Kaplan, S. N., and P. Strömberg (2003), "Financial contracting theory meets the real world: An empirical analysis of Venture Capital contracts," *The Review of Economic Studies,* 70 (2): 281–315.

— (2004), "Characteristics, contracts, and actions: Evidence from Venture Capitalist analyses," *The Journal of Finance,* 59 (5): 2177–2210.

Kumar, P.; Langberg, N., and D. Zvilichovsky (2016), (Crowd)Funding innovation: Financing constraints, price discrimination and welfare, *SSRN Scholarly Paper ID* 2600923, Social Science Research Network, Rochester, NY.

Kuppuswamy, V., and B. L. Bayus (2017), "Does my contribution to your crowd funding project matter?," *Journal of Business Venturing,* 32: 72–89.

— (2018), "Crowdfunding creative ideas: The dynamics of project backers in Kickstarter," in *The Economics of Crowdfunding:* 151–182, Springer.

Li, E., and J. S. Martin (2016), "Capital formation and financial intermediation: The role of entrepreneur reputation formation," *Journal of Corporate Finance* (in press).

McKenny, A. F.; Allison, T. H.; Ketchen, D. J.; Short, J. C., and R. D. Ireland (2017), "How should crowdfunding research evolve? A survey of the Entrepreneurship Theory and Practice Editorial Board," *Entrepreneurship Theory and Practice,* 41 (2): 291–304.

Mollick, E. R. (2013), Swept away by the crowd? Crowdfunding, venture capital, and the selection of entrepreneurs, *SSRN Scholarly Paper ID* 2239204, Social Science Research Network.

Mollick, E. (2014), "The dynamics of crowdfunding: An exploratory study," *Journal of Business Venturing,* 29 (1): 1–16.

— (2016), Containing multitudes: The many impacts of Kickstarter funding, *SSRN Scholarly Paper ID* 2808000, Social Science Research Network.

Mollick, E. R., and V. Kuppuswamy (2014), After the campaign: Outcomes of crowdfunding, *Technical Report* 2376997, UNC Kenan-Flagler Research Paper.

Mollick, E., and R. Nanda (2015), "Wisdom or madness? Comparing crowds with expert evaluation in funding the arts," *Management Science,* 62 (6): 1533–1553.

Moritz, A., and J. H. Block (2016), "Crowdfunding: A literature review and research directions," in *Crowdfunding in Europe:* 25–53, Springer.

OECD (2011), *Financing High-Growth Firms: The Role of Angel Investors,* OECD Publishing.

Ordanini, A.; Miceli, L.; Pizzetti, M., and A. Parasuraman (2011), "Crowdfunding: Transforming customers into investors through innovative service platforms," *Journal of Service Management,* 22 (4): 443–470.

Palfrey, T. R., and H. Rosenthal (1984), "Participation and the provision of discrete public goods: A strategic analysis," *Journal of Public Economics,* 24 (2): 171–193.

PwC/CB Insights (2018), Moneytree Report.

Rau, R.; Gray, M.; Westerlind, L.; Burton, J.; Cogan, D., and A. Lui (2017), The Africa and Middle East alternative finance benchmarking report, Cambridge Center for Alternative Finance.

Rin, M. D.; Hellmann, T., and M. Puri (2013), "A survey of Venture Capital research," in G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance,* Volume 2: 573–648, Elsevier.

Rochet, J.-C., and J. Tirole (2006), "Two-sided markets: A progress report," *The Rand Journal of Economics,* 37 (3): 645–667.

Sayedi, A., and M. Baghaie (2017), Crowdfunding as a marketing tool, *SSRN Scholarly Paper ID* 2938183, Social Science Research Network, Rochester, NY.

Schwienbacher, A. (2017), "Entrepreneurial risk-taking in crowdfunding campaigns," *Small Business Economics:* 1–17.

Shane, S. (2005), Angel Investing: A report prepared for the Federal Reserve Banks of Atlanta, Cleveland, Kansas City, Philadelphia and Richmond, *SSRN Scholarly Paper ID* 1142687, Social Science Research Network, Rochester, NY.

— (2012), "The importance of Angel Investing in financing the growth of entrepreneurial ventures," *Quarterly Journal of Finance,* 2 (2).

Shane, S., and D. Cable (2002), "Network ties, reputation, and the financing of new ventures," *Management Science,* 48 (3): 364–381.

Short, J. C.; Ketchen, D. J.; McKenny, A. F.; Allison, T. H., and R. D. Ireland (2017), "Research on crowdfunding: Reviewing the (very recent) past and celebrating the present," *Entrepreneurship Theory and Practice,* 41 (2): 149–160.

Skirnevskiy, V.; Bendig, and M. Brettel (2017), "The influence of internal social capital on serial creators' success in crowdfunding," *Entrepreneurship Theory and Practice,* 41: 209–236.

Sohl, J. E. (2017), A cautious restructuring of the Angel market in 2016 with a robust appetite for seed and start-up investing, *Technical report,* Center for Venture Research.

Stanko, M. A., and D. H. Henard (2017), "Toward a better understanding of crowdfunding, openness and the consequences for innovation," *Research Policy,* 46: 784–798.

Steigenberger, N. (2017), "Why supporters contribute to reward-based crowdfunding," *International Journal of Entrepreneurial Behaviour and Research,* 23: 336– 353.

Strausz, R. (2017), "A theory of crowdfunding: A mechanism design approach with demand uncertainty and moral hazard," *American Economic Review,* 107 (6): 1430–1476.

U. S. GAO (2000), Small business: Efforts to facilitate equity capital formation, *Technical Report* GGD-00-190, U. S. Government Accountability Office.

Vismara, S. (2015) Information cascades among investors in equity crowdfunding, *SSRN Scholarly Paper ID* 2589619, Social Science Research Network, Rochester, NY.

— (2016), "Equity retention and social network theory in equity crowdfunding," *Small Business Economics,* 46: 579–590.

Vulkan, N.; Åstebro, T., and M. F. Sierra (2016), "Equity crowdfunding: A new phenomena," *Journal of Business Venturing Insights,* 5: 37–49.

Wallmeroth, J.; Wirtz, P., and A. P. Groh (2018), Venture Capital, Angel Financing, and crowdfunding of entrepreneurial ventures: A literature review, *Foundations and Trends in Entrepreneurship,* 14 (1): 1–129.

Wardrop, R.; Rosenberg, R.; Zhang, B.; Ziegler, T.; Squire, R., and J. Burton (2016), Breaking new ground, *The Americas Alternative Finance Industry Report* 1, Cambridge Center for Alternative Finance.

Welch, I. (1992), "Sequential sales, learning, and cascades," *The Journal of Finance,* 47 (2): 695–732.

Wong, A.; Bhatia, M., and Z. Freeman (2009), "Angel finance: the other Venture Capital," *Strategic Change,* 18 (7/8): 221–230.

Wroldsen, J. (2016), "Crowdfunding investment contracts," *Virginia Law and Business Review,* 11: 543.

Zhang, B. Z.; Ziegler, T.; Garvey, K.; Ridler, S.; Burton, J., and N. Yerolemou (2017), Entrenching innovation, *The UK Alternative Finance Industry Report,* 4, Cambridge Center for Alternative Finance.

Zheng, H.; Li, D.; Wu, J., and Y. Xu (2014), "The role of multidimensional social capital in crowdfunding: A comparative study in China and US," *Information & Management,* 51 (4): 488–496.

Ziegler, T.; Reedy, E. J.; Le, A.; Kroszner, R. S.; Zhang, B., and K. Garvey (2017), Hitting stride, *The Americas Alternative Finance Industry Report,* Cambridge Center for Alternative Finance.

Ziegler, T.; Shneor, R.; Garvey, K.; Wenzlaff, K.; Yerolemou, N.; Hao, R., and B. Zhang (2018), Expanding horizons, *Europe Alternative Finance Industry Report,* 3, Cambridge Center for Alternative Finance.

DIGITIZATION AND THE CONTENT INDUSTRIES[1]

Luis AGUIAR

Joel WALDFOGEL

Abstract

Over the last decade, digitization has drastically affected the content industries and the way creative products are consumed, produced, and distributed. This chapter presents empirical evidence on the effects of digitization on revenues, production, and welfare, focusing more specifically on the market for music. We discuss how technological change —despite leading to significant decreases in revenues— enabled an increase in the creation of new products, leading to substantial welfare benefits. We then turn to the evidence regarding the new business opportunities enabled by digitization and discuss how new distribution and consumption platforms like Spotify affected sales and revenues in the music industry. We finally discuss how the global nature of these platforms can affect overall consumption and production patterns in the music and the movie markets around the globe.

Key words: Digitization, content industries.

JEL classification: L82, O33.

[1] The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

I. INTRODUCTION

By drastically reducing the costs of copying and disseminating information, digitization dramatically affected many of the content industries, allowing individuals to experience a radical increase in their ability to consume cultural products such as music, books, or movies. Following the advent of file-sharing networks, much of this consumption was based on copyright infringing content, and the first effects of digitization were painfully experienced by the recorded music industry whose ability to generate revenues was importantly challenged by rampant piracy. With its global revenues plummeting since the birth of Napster in 1999, the recorded music sector –together with other content industries– saw a threat in the advent of digitization, mainly due to its negative effects on revenue and the ensuing implications regarding investment in content.

Despite their negative effects on producers' revenues, digital technologies have also reduced the costs of production, distribution, and promotion of media content, drastically lowering the costs of bringing new products to market (Waldfogel, 2013). The net effect of these two opposing forces –a negative effect on both revenues and costs– has resulted to be positive, leading to an important increase in the number of newly released creative products since digitization. Because product quality is often unpredictable, this increase in the proliferation of cultural products has led to a significant increase in the appeal of newly released products and to substantial welfare benefits.

Digitization has moreover brought forth new business models, generating new opportunities to increase revenues. Online streaming services have exploded in popularity in the past years and are importantly affecting individuals' consumption patterns and revenue sources in the music industry. While the evidence indicates that they are depressing sales of recorded music, the important increase in music streaming services is now bringing total recorded music revenues back to growth in certain countries. In the US, revenues from streaming services grew 68.5% to $3.9 billion from 2015 to 2016, leading recorded music revenues to increase 11.4% to a total of $7.65 billion over the same period.

Beyond their direct effects on revenues, the emergence of streaming platforms in the music and the movie industries has importantly affected the patterns of trade in cultural products. Digital retailing has offered producers and creators the possibility of reaching much larger markets than ever before. For consumers, digital distribution platforms offer new opportunities to access and discover new products, both domestic and foreign. By increasing the set of works available around the world, digitization effectively decreased the cost of trade and consequently created new opportunities for cultural exchange.

This chapter presents and discusses the empirical evidence regarding the effects of digitization on the content industries. With a specific focus on the market for music, we begin by showing how digitization –despite leading to significant decreases in revenues– enabled an increase in the creation of new works, leading to substantial welfare benefits. We then turn our attention to the new business opportunities enabled by digitization. More specifically, we discuss how new distribution and consumption platforms like Spotify affected sales and revenues in the music industry. We finally discuss how the global nature of these platforms can affect overall consumption and production patterns in the music and the movie markets around the globe.

II. THE EFFECTS OF DIGITIZATION ON REVENUES

The recorded music industry offers a striking example of the potential negative effects of digitization on producers' revenues. In the blink of an eye, the advent of file-sharing networks offered consumers around the world the opportunity to access a vast amount of music without having to compensate right-holders in any form. Following the appearance of Napster in 1999, global recorded music revenues took a dramatic nosedive, which would continue for years.[2] Industry observers have naturally and understandably seen this sustained reduction in revenues as a serious concern. Because of the large investments needed to bring creative products to market, a reduction in revenues can lead to a reduction in creative output, ultimately hurting both consumers and producers. This concern consequently generated a great debate among academics, who have for many years sought to identify the effects of piracy on music sales.

The effect of piracy on recorded music revenue is ambiguous as it depends on the types of consumers that decide to consume without paying. While some individuals may value a product (*e.g.,* a song or album) positively, their valuation may still be below the market price. Because these individuals would by definition never purchase the product in question, acquiring it via unpaid means would not affect producers' revenue. In fact, piracy would turn the deadweight loss initially experienced by these individuals into consumer surplus. On the other hand, some other consumers may value the product above the market price and nevertheless decide to consume it without paying. In that case, piracy would naturally decrease revenues as these instances of unpaid consumption would have been converted into sales absent piracy.[3]

---

[2] See, for instance, the 2017 global music report from the International Federation of the Phonographic Industry (IFPI), available at http://www.ifpi.org/downloads/GMR2017.pdf

[3] Because all instances of sales displacement will convert producers' revenues into consumers surplus, unpaid consumption would unambiguously increase welfare in the short run. Decreases in revenues could, however, affect the production of new music if producers can no longer cover their costs. Piracy could therefore destroy all the surplus in the long run. See Section III.

Measuring the effect of piracy on recorded music sales is an inherently difficult task, mainly for two reasons. First, piracy is an illegal behaviour, which renders its measurement difficult. It is therefore challenging to link data on purchases with data on piracy, let alone to obtain data on volumes of unpaid consumption. Second, even if such data on volumes of both sales and unpaid consumption is available, identifying the causal effect of music piracy on sales is extremely challenging given the non-experimental nature of the data. Because music sales and piracy are both driven by the unobserved popularity of music, piracy is itself an endogenous variable. This would therefore result in a positive correlation between piracy and sales even if piracy does not cause an increase in purchases.

Researchers have pursued several empirical approaches to analyze the effect of music piracy on sales. A first approach relies on individual-level data, asking whether consumers who engage in piracy engage in more or less paid consumption. A second approach uses product level data to see whether records that are pirated more are purchased more or less (Oberholzer-Gee and Strumpf, 2007). Both approaches naturally suffer from the endogeneity issue mentioned above, and researchers have relied on various empirical strategies to identify the causal effect of piracy on sales.[4] Some have used access to broadband and Internet connection speed as sources of exogenous variation in piracy (Rob and Waldfogel, 2006; Zentner, 2006). Others have relied on geographical variation in piracy levels (typically proxied by measures of Internet broadband penetration), asking whether places with higher piracy levels have lower levels of sales (Hui and Png, 2003; Peitz and Waelbroeck, 2004; Liebowitz, 2008).

After more than a decade of research, most of the evidence indicates that piracy has indeed depressed recorded music sales.[5] The estimates from most studies show that the impact of piracy was large enough to have caused most if not all of the decline in record sales since the advent of Napster (Liebowitz, 2016).[6]

---

[4] See Smith and Telang (2012) for an overview of the different approaches used in the literature to identify the effects of piracy on recorded music sales.

[5] Studies focusing on the effects of piracy on movie and book sales also find significant displacement effects (Rob and Waldfogel, 2007; Bai and Waldfogel, 2012; Reimers, 2016).

[6] A large body of empirical literature has consequently focused on the effectiveness and consequences of copyright enforcement efforts in the music industry (Danaher *et al.,* 2014; Adermon and Liang, 2014), the movie industry (Danaher and Smith, 2014; Penkert, Claussen and Kretschmer, 2017; Aguiar, Claussen and Peukert, 2018), and the book publishing industry (Reimers, 2016), among others. See Danaher, Smith and Telang (2013) for an overview of copyright enforcement mechanisms in the creative industries.

III. THE EFFECTS OF DIGITIZATION ON PRODUCTION AND WELFARE

With the objective of providing creators incentives to bring new products to market, copyright grants creators monopoly rights over their works. This allows them to generate revenues from selling their output, and therefore recoup their initial investments. While the impact of piracy on revenue is an important question for sellers of recorded music, it is not the only question of interest for welfare and public policy. Piracy may undermine revenues –and therefore potentially undermine investments– but assessing the good functioning of creative industries in the digital era consists in asking whether creators still bring forth valuable new products.

Why would creators not reduce their production following a reduction in their revenues? It is crucial to highlight that the implications of digitization go beyond their effects on the ability of consumers to copy and share existing content. Technological change has also allowed major improvements in the production process of many different industries. In the case of the music industry, it has now become very easy to produce, distribute, and promote new music to consumers around the world. Production of recorded music has become less costly as relatively inexpensive computers and software are now able to perform the functions of costly studio equipment. Digital distribution has made it possible for artists' works to be available to millions of consumers without the costs of pressing discs, transporting physical goods, or maintaining inventory in physical retail stores. Finally, promotion has become less expensive as Internet radio, social media, and widely available online criticism have supplemented the traditional promotional bottleneck of terrestrial radio.[7]

Digitization has consequently had two opposing effects on the incentives to bring new products to market. On the one hand, technological change has depressed creators' ability to generate revenue. On the other hand, it has allowed them to create products at a lower cost. In light of the important increase in the number of products observed in the creative industries, the net effect of these two opposing forces clearly seems to have been positive. As we will discuss in more detail below, the music industry has indeed witnessed an important increase in the number of newly released titles (Oberholzer-Gee and Strumpf, 2010; Handke, 2012; Aguiar and Woldfogel, 2016). Similar increases have been documented in other industries such as movies (Waldfogel, 2016) and books (Waldfogel and Reimers, 2015).

[7] See Waldfogel (2013) for a discussion of the cost reductions in the music industry. Technological change has allowed for similarly important costs reductions in other industries such as movies and books. The reader is directed to (Waldfogel, 2017) for a detailed and insightful account of these changes and their consequences in alternative creative industries.

The increase in the availability of new content faced by consumers following digitization has typically been described and analyzed through the "long-tail" phenomenon.[8] Prior to the advent of the Internet, the choice set faced by consumer was limited to the shelf-space available in brick-and-mortar stores. By liberating retailers from physical restrictions, digitization allowed consumers to access and benefit from a much larger set of products online. As an example, consider the full set of books available at Amazon –an online retailer with infinite shelf space– compared to, say, the 100,000 titles available in local book stores. Accessing this larger set of books would naturally bring important benefits to consumers. Although each of the additional books available online has lower demand than the ones in stores –justifying their non-availability in a purely physical environment– the additional benefits brought by many additional books can lead to large welfare increases.

1. Quality Unpredictability and the Welfare Benefits of New Products

While online retailing has clearly benefited consumers by allowing sellers to carry a much larger choice of products (the long-tail), the lowering of costs brought forth by digitization has also allowed for a large entry of new products by creators, including those that would not have been released in a pre-digitized world. Because the quality of many products is unpredictable at the time of investment –which is particularly true in creative industries– the welfare benefits of bringing these new products to market can be substantial.

The mechanism behind this idea is fairly simple. Consider the introduction of new products whose commercial success is perfectly known at the time of investment. Assume, for instance, that record labels are able to perfectly assess the revenue Y that each title or artist would earn if it were released. Assume further that the record label would need to incur a cost equal to C in order to release the product. In that case, all projects with $Y>C$ would be released. Following technological change, the cost of releasing products falls from C to C', and more products can be released. By construction, all of these newly released products would have lower commercial success than the ones already in the market. When commercial success is perfectly predictable, a decrease in entry costs therefore leads producers to effectively add less popular products to the market, and the welfare benefits from these products result uniquely from the additional and infinite shelf-space provided by the Internet.

When commercial success is hard to predict at the time of investment, record labels will have to form a guess about the success of a new release. In

[8] See Anderson (2006) for a popular account of the long tail.

other words, they will forecast the commercial appeal of a title or artist as their true appeal Y plus an error term ε: $Y'=Y+\varepsilon$. Their release threshold will now depend on their *expected* appeal Y', and they will decide to release a product whenever $Y'>C$. Note, however, that there will be some products characterized by $Y>C$ and $Y'<C$: products that *should* be released –because their *realized* commercial success is larger than their cost– but won't be released because their expected success is too low. In other words, some products would be successful if released, but the unpredictability surrounding their commercial appeal prevents them from being released in the first place. By reducing entry costs from C to C', digitization allows for the release of some of these particular products. The benefits of digitization should therefore not only be seen from the perspective of an extension of shelf-space, but they should also account for the fact that lower entry costs enable the creation of appealing products that would otherwise not have been released.

Unpredictability of commercial success is a common feature of the creative industries, and we should therefore expect substantial welfare benefits from an increase in new products following digitization. Industry observers report that roughly 10 percent of new movies are commercially successful, with similar figures for music and books (Caves, 2000; Vogel, 2014).[9]

In a context where commercial success is hard to predict —such as the music industry— an increase in product entry would have several empirically testable implications. First, the average appeal or quality of newly released products would increase.[10] Second, a significant number of products that were expected to fail –and were therefore not brought to market in a pre-digitized world with higher entry costs– would account for a growing share of the successful products. Finally, a growth in entry would not necessarily reduce sales concentration since new products could also attract substantial sales. We now turn to the empirical analysis of these implications in the music industry.

2. The Evolution of Product Quality

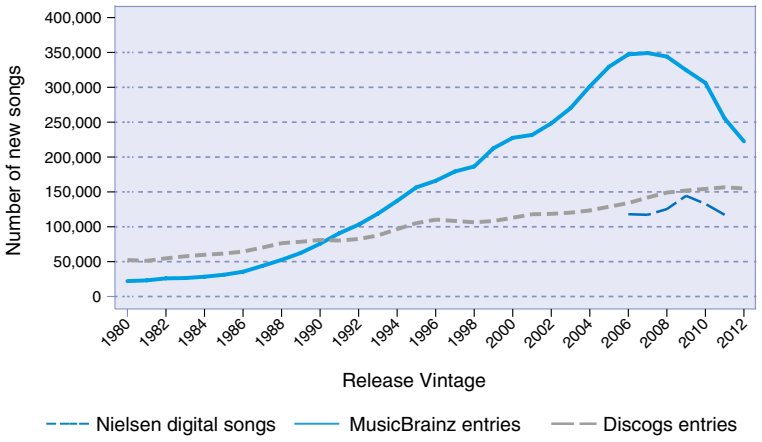
As noted above, the music industry has witnessed an important growth in the number of newly released products over the last decade. Figure 1 presents the number of new recorded music titles released over time, from three different sources. While the figures differ across sources, the numbers clearly indicate a

[9] Screenwriter William Goldman also famously remarked that nobody knows anything about which movie releases will be appealing to consumers. See Goldman (1989).

[10] Despite the creative context, the word "quality" has no aesthetic connotations and is only used to denote the service flow implied by consumption decisions.

FIGURE 1

NUMBER OF NEW SONGS BROUGHT TO MARKET

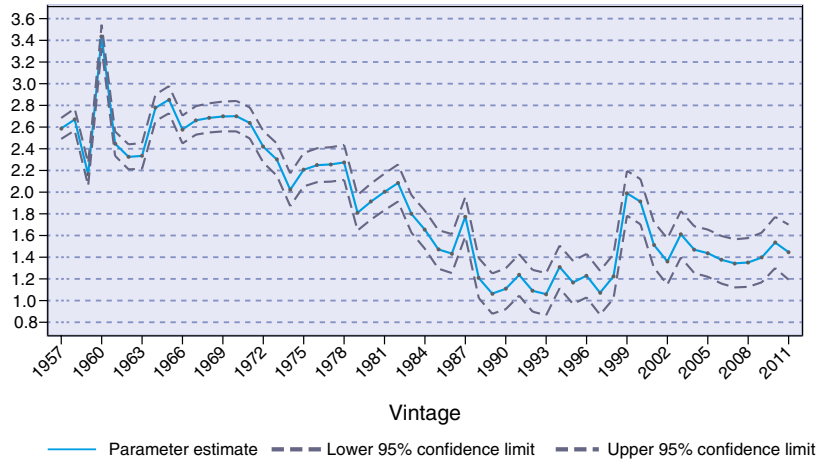


Source: Aguiar and Waldfogel (2016), Figure 2.

sharp increase in the number of new works produced per year, and in particular since 2000. Additional reports indicate that the number of new music products brought to market tripled between 2000 and 2008 (see Aguiar and Waldfogel, 2018b).

Observing a large increase in the number of works created since digitization naturally does not imply an increase in the quality of these new products. It is possible, after all, that creators are releasing works that are not necessarily appealing to consumers. While measuring music appeal is naturally a challenging task, Waldfogel (2012) proposes a method for inferring its evolution over time. By relying on consumers' purchase decisions, one can ask whether particular vintages of music are used more intensively than others, after accounting for the fact that older music vintages tend to be used less than more recent ones. Relying on both purchase data as well as critics' judgment for the US, he finds that music's appeal –in the eyes of US consumers– grew sharply since 1999. Aguiar and Waldfogel (2016) perform a similar exercise by relying on digital music sales data on 17 countries over the period 2006-2011. By observing sales of different music vintages in multiple calendar years, they construct a similar index of the appeal of each particular vintage. Figure 2 shows the evolution of this quality index. The index jumps in 1999-2000, indicating that the vintages of music released since 1999 are more used

FIGURE 2
INDEX OF MUSIC QUALITY



Source: Aguiar and Waldfogel (2016), Figure 4.

than previous vintages. In other words, relative to the vintages from the 1990s, the appeal of music released after 1999 increased in the eyes of consumers of these 17 countries.[11]

If an increase in music quality results from digitization through the mechanism described above, we should also observe a growing share of products with low *ex ante* appeal within the set of successful products. These are the products that were expected to perform poorly –and were therefore not released when entry costs were high– but ended up being successful following their release. Using the notation above, these are the products for which $C' < Y' < C$, and releasing these products would lead to an increase in music appeal.

Identifying products with low *ex ante* promise –the artists that were not deemed "good enough" to be released prior to digitization– is inherently difficult. Waldfogel (2012 and 2015) and Aguiar and Woldfogel (2016) consider independent labels releases as products with low expected commercial success,

[11] The countries included in the sample are: Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland,the United Kingdom, the United States, and Canada. Analyzing the evolution of music quality for each of these countries leads to similar results, indicating that quality has increased in the eyes of consumers in each of these 17 locations.

assuming that they were not promising enough to have been signed by major labels prior to digitization.[12] Waldfogel (2015) shows that the share of top-selling albums released by independent labels increased from 12% to 35% between 2000 and 2010. Since these new products attract a substantial share of sales Aguiar and Waldfogel (2016), also observe an increase in sales concentration. These results indicate that an important share of valuable products would not have been created and been made available for consumers had digitization not occurred.

These patterns have also been documented in other cultural industries heavily affected by digitization and characterized by a high unpredictability of products' appeal. In the market for books, Waldfogel and Reimers (2015) show that the share of best-selling books that were originally self-published grew from 0 to over 10 percent between the appearance of the Kindle in 2007 and 2014. This share increased to as much as 40% in the romance category. In the case of the movie industry, Waldfogel (2016) finds that the share of box office and DVD revenue accounted for by independent films increased from 20 to around 40 percent.

3. Quantifying the Welfare Benefits from New Products

Digitization has clearly benefited consumers by allowing them to access a larger set of products. As mentioned above, most of the literature quantifying the welfare benefits of digitization focused on the "long-tail" phenomenon, measuring the benefits of having access to an unrestricted shelf-space of products via online retail. It is clear that larger online choice sets have importantly benefited consumers compared to their limited offline counterpart.[13] For instance, Brynjolfsson, Hu and Smith (2003) estimate that the benefit consumers obtain from accessing a long tail of additional varieties of books may be as high as $1.03 billion per year in 2000.

The evidence presented above indicates that digitization also affected welfare in a different way. By decreasing the costs of bringing new products to market, digitization has enabled entry of a large set of new products. Because the quality of these products is difficult to predict at the time of release, many

[12] There has been a substantial growth in both the number of products coming to market and the labels bringing these products (Oberholzer-Gee and Strumpf, 2010; Handke, 2012). See Waldfogel (2017) for a more detailed discussion related to the identification of products with low *ex ante* appeal in the music, book publishing, and movie industries.

[13] Quan and Williams (2017) highlight that one should take into account the fact that offline products are also tailored to local tastes. Overlooking this fact would therefore overstate the benefits of having access to a larger choice set.

of them end up being highly appealing to consumers. Following the mechanism described in Section III.1, digitization therefore not only offered consumers a larger choice set; it also changed its composition to include highly valuable new products that would not have been created absent the decrease in entry costs.

Aguiar and Waldfogel (2018b) quantify the size of the welfare benefits of new music releases and evaluate how these benefits are affected by the unpredictability that characterize creative products. More specifically, they quantify the increase in consumers' welfare following a tripling in the number of new music releases. When music quality is perfectly predictable, a cost reduction that enables a tripling in the number of new releases will bring consumers limited benefits as the new songs will be –by construction– less appealing than the songs that were already available to them. Put in a different way, this would lead consumers to have access to a larger set of low-appeal products. When music quality is unpredictable, however, an increase in the number of new releases will include songs that were initially not deemed good enough to have been released prior to the reduction in costs. But because of unpredictability, some of these will still end up being highly appealing to consumers. Aguiar and Waldfogel (2018b) estimate that a tripling in the number of new releases according to expected appeal adds about 20 times more benefit to consumers than a tripling of products according to realized appeal. In other words, the market introduction of highly valuable new products following digitization —many of which would not have been created absent the decrease in entry costs due to their lower expected appeal— has had large welfare benefits compared to the conventional benefits of the long-tail. Because the commercial success of new products is also unpredictable in many other industries, this idea may also be applicable outside of the creative industries.[14]

IV. DIGITIZATION AND NEW BUSINESS MODELS

On top of its effects on new product entry, the digitization of the media industries has also brought forth many new business model opportunities, potentially holding the promise of helping to increase revenues. In the music industry, online music streaming services –which essentially allow consumers to listen to music without the need to download the corresponding audio file– have importantly expanded music consumption opportunities by offering consumers access to large bundles of music.

[14] For instance, Gourville (2005) documents new product failure rates of between 40 and 90 percent across many categories.

1. Streaming in the Music Industry

Streaming music services have exploded in popularity in the past few years, leading to both optimism and concern about their impacts on recorded music revenue. On the one hand, these large bundles of products hold –at least in principle– the opportunity to raise revenue by exploiting the zero marginal costs of production characterizing digital formatting as well as differences in the willingness to pay of consumers.[15] A consumer subscribing to a streaming service will generate revenue for each song they listen to, including the ones they value too little to have purchased a la carte (through platforms like iTunes for instance). Streaming therefore offers the possibility of converting some willingness to pay into revenue for individuals who would have forgone these instances of consumption in a world where only a la carte options are available. Likewise, streaming may increase revenue by turning individuals who consume music via unlicensed channels (piracy) into paying customers.

On the other hand, and because streaming serves as a new form of music consumption, these platforms are also likely to directly affect other sources of revenues such as recorded music sales. Determining whether streaming stimulates or displaces the sales of recorded music is therefore crucial to understand its impact on the recorded music industry. If streaming serves as a promotional tool that can stimulate demand through other channels, then these platforms would unambiguously raise recorded music revenue. If streaming serves as a substitute for recorded music sales, its effects on revenues would depend on the rate at which displacement occurs.

To make matters more complicated, specific functionalities characterizing streaming services may further affect the way they interact with alternative consumption channels. Streaming platforms can broadly be divided into two distinct categories: *interactive* and *non-interactive platforms*. The non-interactive platforms –such as Pandora in the US– offer services that are similar to a radio broadcast in that the end user is offered a pre-programmed set of songs. While this selection of songs is usually based on algorithms that take musical preferences into account, consumers cannot select the songs they want to listen to or even observe the order of the tracks to be played. This is in contrast with interactive platforms –such as Spotify, Apple Music, or Deezer– which offer consumers the liberty to pick the songs they want to listen to, provided these are available in the platform's repertoire.

[15] See the literature on bundling in general (Admas and Yellen, 1976; Schmalensee, 1984), the bundling of information goods (Bakos and Brynjolfsson, 1999), and music bundling in particular (Shiller and Waldfogel, 2011).

By providing users with on-demand access to songs almost anywhere, interactive services appear to be an appealing alternative to buying music and are therefore likely to serve as strong substitutes for music purchases. Conversely, non-interactive services only propose songs and artists to consumers based on their musical preferences. Because users are not allowed to pick what songs they want to consume, these services can expose individuals to songs they would otherwise not have heard and can therefore act as discovery tools that can ultimately stimulate music demand. It is therefore easy to imagine that non-interactive services could act as a complement to music purchasing.

On top of the above distinction, interactive streaming platforms usually offer two types of services. The first one comes in the form of a premium subscription, which typically provides users with on-demand, advertisement-free listening on fixed and mobile devices, both online and offline for a monthly fixed-fee. The second type of service is free and supported by advertisement. While it usually offers unlimited access to streaming, this free service is typically provided with further restrictions, particularly with respect to the mobility of access. In particular, interactive streaming mobility is drastically restricted for free users who are typically imposed the use of shuffle mode on mobile, have no ability to skip tracks within playlists, or have their repeated listening restricted. In other words –and as opposed to premium accounts– free subscription are preventing users from flexibly accessing music everywhere. Assuming that users positively value mobility in their music listening –and if streaming indeed can be used as a product discovery tool– restrictions in mobile streaming may lead free users to complement their streaming with the purchasing of music. For instance, a user who discovers a new song through a free streaming account may decide to further purchase it –or pirate it– in order to access it unrestrictedly through a mobile device. From that perspective, the effects of free and mobile-restricted interactive streaming are potentially similar to the effects of non-interactive streaming.

2. The Effects of Music Streaming

Regardless of the type of service considered, identifying the effect of streaming on music sales is an inherently challenging task. Comparing levels of streaming and sales for a particular song will naturally suffer from the fact that streaming and sales may both be driven by the song's popularity. This would result in a positive correlation between streaming and sales even if streaming does not cause an increase in purchases.[16]

[16] This issue is similar to the one that plagues the identification of the effect of music piracy on sales and which was discussed in Section II.

In order to identify the effect of non-interactive streaming on sales, data scientists at Pandora have undertaken experiments where they would stop the plays of certain songs in randomly selected geographic areas to see what happened to the sales of these same songs. McBride (2014) finds that Pandora increases music sales by around 2%, providing evidence that non-interactive music streaming services can stimulate sales. In a similar vein, Danaher (2014) uses data from an Internet consumer panel tracking company and shows that the use of non-interactive webcasting services indeed has a significantly more positive impact on digital song purchases than interactive webcasting services.

Aguiar (2017) relies on the introduction of a listening cap on free streaming by the French leading streaming platform Deezer to explore how free and mobile-restricted interactive streaming affect alternative sources of digital music consumption such as digital purchases or piracy. Using Internet clickstream data, which allow to precisely follow the online behavior –including visits to licensed and unlicensed digital music consumption websites– of a representative sample of 5,000 French Internet users during the year 2011, the results show a negative effect of the imposition of the free streaming cap on visits to both licensed and unlicensed music downloading websites. These findings therefore indicate a positive effect of free and mobile-restricted streaming on these alternative sources of consumption.

Datta, Knox and Bronnenberg (2017) analyze the effects of Spotify adoption on individual music consumption and discovery. While they do not distinguish between free and premium subscriptions, they find that the adoption of Spotify increases overall music consumption but cannibalizes consumption on iTunes. Their results also show that adopting Spotify leads to an increase in the variety of music consumed and to more discovery of music. Wlömert and Papies (2015) is one of the few papers to distinguish between the effect of free and premium streaming adoption on online recorded music purchases. Relying on a survey panel of music consumers, they find that consumers who adopt such services purchase significantly less recorded music, with a larger cannibalization effect for paid streaming adoption.

Another important challenge related to the identification of the overall effect of streaming on sales is what Liebowitz (2004) calls a "fallacy of composition." Even if we could properly identify the causal effect of streaming on sales *at the song level,* the latter would still not be helpful to identify *the overall* effect of streaming on sales. To understand this point, suppose that streaming does increase the relative popularity of individual songs, and is therefore capable of increasing sales of more heavily streamed songs. It would still be possible that the emergence of streaming platforms leads some consumers to stop purchasing songs. The fact that streaming could stimulate a particular song's

sales is therefore not relevant to answer the question of whether streaming reduces sales overall.

One way of analyzing the overall effect of streaming on the recorded music sales is to concentrate on a time period in which streaming increased significantly. Measuring the effects of this growth on aggregated recorded music sales would then provide evidence on whether streaming displaces or stimulates purchases overall. Aguiar and Waldfogel (2018a) take advantage of the important growth in streaming during the years 2013-2015 to measure its collective impact on the sales of recorded music. By relying on aggregate US data on weekly digital and physical music sales, they find a negative and statistically significant relationship between volumes of streaming and sales.[17] More specifically, their estimates indicate that an additional thousand collective streams reduces track-equivalent sales by between 1.44 and 2.85.[18]

Because every instance of streaming also generates revenue to right holders, identifying the rate of sales displacement is not enough to evaluate the effect of streaming on revenues. More specifically, one needs to assess whether the loss in revenues from displaced sales is offset by the payments obtained through streaming. Total revenue to right holders consists in the sum of revenue from music sales and from music streams. Revenues from music sales in turn correspond to the number of tracks sold times the payment per track accruing to right holders. Revenues from streaming correspond to the total number of streams times the payment per stream. Evaluating the overall effect of streaming on revenues therefore also requires information on the payment that right holders obtain for each sale, and the corresponding payment obtained for each additional stream.[19]

The payment accruing to right holders for each additional sale can be calculated in a relatively straightforward way, at least for digital downloads. Right holders roughly receive 70 percent of the revenue from digital track sales, and given that the average revenue per digital track sold was $1.174 in 2014, this leaves $0.822 per track sold to be shared among the right holders. Based on these figures, an additional 1,000 streams reduce revenues from track-equivalent sales by between $2.34 and $1.18.[20] What about streaming payments? These

[17] They also identify an negative and significant effect of streaming through Spotify on piracy using data on 18 European countries, the US, Canada, Australia, and New Zealand during 2012-2013.

[18] Track-equivalent sales are defined as digital track sales plus (10×album sales).

[19] The right holders of recorded music are some combination of record labels, musical performers, song writers, and music publishers. These can be different entities, for instance when a performing artist releases a song written by another person on an album released by a major label. They can also be the same entities, for instance when a self-released artist performs their own composition.

[20] Note that $2.85 \times \$0.822 = \$2.34$ and $1.44 \times \$0.822 = \$1.18$.

are substantially harder to obtain, but relying on public sources, payments are estimated to vary from $1.51 to $2.77 per thousand streams.[21]

To sum up –and based on the above figures– an additional thousand streams consequently raise revenues by between $1.51 and $2.77, while they reduce track-equivalent sales revenue by between $2.34 and $1.18. The overall effect of music streaming on right holders' revenue therefore heavily depends on the per-stream payments. For values of streaming payments at the higher end of the range, streaming appears to increase overall revenue. If streaming payments are at the lower end of the range –and are therefore less effective in offsetting the sales displacement effect– then one cannot reject the idea that streaming has been revenue neutral as of 2015.

While streaming seems to have been mostly revenue-neutral for the recorded music industry as of 2015, the transition from sales to streams continued in the years following the above study period. More recent US aggregate statistics provide evidence of an important displacement of sales by streaming –interactive and paid streaming in particular– in 2016. According to the Recording Industry Association of America (RIAA), US recorded music revenue reached $7.65 billion in 2016, a 11.4% increase relative to 2015. This increase was driven by a 68.5% growth in streaming revenue, and more particularly by a large growth in paid on-demand streaming, which grew 95% to $2.3 billion. Digital downloads revenue decreased by 21.6% over the same time period.[22] More recent data shows that paid subscriptions to streaming services accounted for 69% of the $2.5 billion generated by streaming during the first half of 2017.[23] Finally, as streaming revenue grew in 2016, so did the average per-stream royalty.[24] These figures indicate that while paid interactive music streaming displaces sales, this new form of consumption is fulfilling its promise of increasing total industry revenue.

Beyond their direct implications for the industry's overall revenue, the effect of music streaming on sales and piracy also has important implications for the various commercial strategies that may be envisioned by rightholders and streaming services themselves. For instance, certain rightholders may decide that their recently released works –which would presumably be in high demand– should initially be left out of streaming platforms and only

[21] See Aguiar and Waldfogel (2018a) for more details on the different types of streaming payments and on these calculations.

[22] See http://www.riaa.com/wp-content/uploads/2017/03/RIAA-2016-Year-End-News-Notes.pdf

[23] See https://www.riaa.com/wp-content/uploads/2017/09/RIAA-Mid-Year-2017-News-and-Notes2.pdf

[24] See https://www.billboard.com/biz/articles/news/digital-and-mobile/7744274/us-music-industry-sees-first-double-digit-growth-in-22

be made available for streaming after an initial increase in permanent sales. From that perspective, streaming services would facilitate inter-temporal price discrimination, potentially increasing revenue for rightholders. A similar windowing strategy could be contemplated across the different tiers of streaming services. One could for instance imagine that new releases be first made available on premium streaming services during periods of high demand and later on free ad-supported tiers. Such strategy may convince some free-tier users of a streaming service to subscribe to a premium plan, potentially generating more revenues for both rightholders and the streaming service. At the same time, such strategies may lead free-tier users to revert to piracy, potentially lowering revenues. The viability of these strategies therefore requires a good understanding of the relationship between streaming, sales, and piracy. Additionally, the success of such approaches may heavily depend on the popularity of the artist in question. One may easily imagine that a popular artist would benefit from a windowing strategy the most, while less popular artists would perhaps rely on streaming services for promotion or discovery. From that perspective, understanding the effects of streaming on the sales and piracy of artists according to their popularity seems like a fruitful avenue for future research.

V. DIGITIZATION AND GLOBAL DISTRIBUTION

Digitization has importantly affected the way music is distributed around the world. Prior to technological change, music producers needed to produce physical products (*e.g.* CDs) and organize distribution through physical record stores near consumers. Distribution of music was therefore costly and the number of products made available to consumers was limited. Because of the high unpredictability of music appeal that we discussed at length above, many of these titles would additionally not necessarily find success.

By eliminating the need for physical products and local retailers, digitization drastically changed the music distribution landscape. Digital formatting led to the emergence of new digital retail platforms —such as the iTunes music store— which largely increased the reach and the number of products made available to consumers. A song available on the iTunes music store would immediately be available to all consumers in the store's country. For artists and music producers, digital distribution therefore offered an opportunity to reach new and potentially larger markets to sell their products. The large increase in music production enabled by digitization, coupled with the benefits of digital retail, therefore led to an important increase in the availability of foreign products within consumers' choice sets in each country. For consumers, digital retail and the unbundling of the music album –which offers the option to purchase

individual songs for about 1€ a piece rather than an entire album– further offered new opportunities to access and discover foreign products more easily. By increasing the set of foreign songs that could be accessed by consumers, digitization effectively decreased the costs of trade, therefore creating more opportunities for cultural exchange.

Many countries express concerns over forces that make cultural products of foreign origins more readily available in their national territory. Foreign products –those from the US and other anglophone countries in particular– are typically seen as a threat to domestic sellers and culture. Local content requirements have consequently been implemented in many places around the world. For instance, Canada, France, Australia, and New Zealand, all regulate the minimum share of domestic content to appear on their domestic radio stations (see Richardson and Wilkie, 2015).

While a reduction in the trade costs of cultural products naturally raises important questions regarding content production and consumption patterns, it is a priori not clear how these outcomes would be affected. Trade is, after all, a two-way street. On the one hand, a greater availability of foreign products could make popular repertoires –such as those of the US and the UK– even more dominant, possibly displacing local cultural production in smaller, non-Anglophone countries. On the other hand, freer trade can increase the availability of products from countries that have not traditionally produced content with sufficient commercial prospects to justify paying the fixed costs of trade.

Despite substantial growth in availability, digital music choice sets have not yet fully converged across countries. This is particularly true for digital music sold through country-specific online retailers, for which cross-border transaction costs are often perceived as an obstacle to greater availability (Gomez and Martens, 2014). But what if choice sets were to fully converge? Would consumption converge towards cultural products from the most popular repertoires? Would local cultural production be displaced, or would smaller-market repertoires be able to benefit from a greater market?

Aguiar and Waldfogel (2014) rely on digital music sales data on 17 countries –the US, Canada, as well as 15 European countries– to simulate and quantify how consumers and producers would benefit from further trade opening in digital music. As it turns out, most of the gains from trade are already realized under the status quo, even if choice sets have not fully converged. In other words, consumers already have access to the products they like the most, and providing them access to an even larger choice set –one that would include all songs available in any country– would only marginally benefit them. Unsurprisingly, consumers who benefit the most from an increase in availability tend to be located in countries with smaller status quo choice sets. Smaller

producers like Finland, Norway, and Sweden –whose titles are less ubiquitously available under the status quo– tend to gain more from a reduction in trade frictions. Their gains from greater market availability seem to offset the losses implied by facing more foreign competition. US products, which are already widely available under the status quo, gain little from frictionless trade. These results therefore indicate that lower trade costs are likely to benefit smaller countries' products the most by allowing them to reach larger markets.

These types of simulations are one way of exploring the effects of a hypothetical world with lower trade costs, one where all existing titles are made available for sales on digital retailing platforms in every country. However, the recent growth in streaming services discussed in Section IV offers another possibility of exploring the effects of a reduction in trade costs following digitization.

1. The Frictionless World of Streaming

In many ways, streaming platforms offer benefits that are similar to the ones provided by other digital retailers like iTunes as consumers can access large catalogs of music, including titles from many foreign repertoires. But streaming offers an additional and important benefit in that the cost of listening to an extra song is zero. Because music is an experience good –meaning that consumers need to listen to a song in order to decide whether they would buy it– this reduction in the cost of experimenting with unknown titles can have important implications.[25] Compared to digital a la carte sales –where the cost of experimenting is fixed at about 1€ per song– streaming further decreases trade frictions by allowing for an easier access to previously unknown products, including foreign ones. From that perspective, streaming services are platforms where product availability is ubiquitous and trade is already frictionless.

To analyze the evolution of music trade patterns since digitization, Waldfogel, Aguiar and Gómez (2017) use data on 17 countries' pop charts and Spotify streaming during the period 2004-2015.[26] They show that trade frictions on the pop charts –measured as the domestic shares of consumption– have declined between 2004 and 2015. In other words, consumers decrease their consumption of domestic music with the increasing access to foreign products. How do these domestic shares compare with the domestic shares on Spotify? If trade frictions were identical for streaming and sales, we should expect similar domestic shares for both channels. What the data show is that domestic shares

[25] Datta, Knox and Bronnenberg (2017) show that taking up streaming services like Spotify can increase discovery of new music.

[26] Rather than quantities sold, the pop charts data provide ranks for the weekly top songs. While theses ranks are mostly based on sales, some are also based on other factors such as airplay or even streaming. In that sense, a comparison of streaming data with pop chart rankings will underestimate the differences in trade patterns based on streaming versus sales.

are lower with streaming in most countries, suggesting that trade frictions are indeed smaller for streaming compared to sales.

On their face, declining domestic shares could be interpreted as bad news for both the production and consumption of local content. One worry may be that products from the most popular repertoires –the US for instance– gain market shares in each country at the expense of domestic products. But domestic shares only tell us part of the trade story. A given repertoire could collect a smaller share of domestic sales following digitization, yet increase its share of world sales. Analyzing the evolution of repertoires' share of world sales shows that while the US origin share of world sales as been declining between 2004 and 2015, the share of European repertoires has been increasing.

To get a glimpse of how lower trade frictions could further affect world market shares, one can compare each repertoire's share of the world market on Spotify relative to sales. Performing this exercise shows that repertoires from a few small countries have larger shares of the world market on Spotify. These include Norway, Austria, Switzerland, Sweden, Portugal, Australia, the Netherlands, the UK, and Canada. The repertoires that do worse on Spotify include larger countries like the US, Germany, and France, among others. These results therefore seem to suggest that digitization may help in leveling the playing field by allowing smaller country producers to reach larger shares of the world market.

A decrease in the costs of trade and the ensuing convergence in choice sets could also lead to convergence in consumption if preferences were identical across countries. To explore whether the patterns of consumption have become more similar across places with digitization, one can characterize the similarity of consumption between two countries according to the origin distribution of the music they consume. For instance, suppose that 75% of the music consumed in country A is domestic and the remaining 25% comes from music produced in country B. In country B, consumers devote 90% of their music consumption to their own productions, and the remaining 10% comes from country A. One can construct a measure of similarity between the consumption patterns of countries A and B by calculating the euclidean distance between their respective consumption vectors.[27] In this example, the distance between countries A and

[27] The distance between the consumption vectors of A and B is is calculated as $\sqrt{\left(Share_A^A - Share_A^B\right)^2 + \left(Share_B^A - Share_B^B\right)^2}$, where $Share_B^A$ is the share of country's A consumption that comes from country B's productions. Note that if both countries consume the same baskets of music, then the distance would be equal to 0. In the opposite extreme case where each country devotes all of its consumption to domestic products only, the distance would be equal to $\sqrt{(1-0)^2 + (0-1)^2} = 1.414$. For each country-pair in the data, this measure of distance would therefore take values between 0 and 1.4.

B would be equal to 0.92.[28] Relying on this distance measure, the data show that the average distance between countries' consumption vectors fell steadily from about 0.5 in 2007 to nearly 0.3 in 2015. Moreover, the average distance based on streaming data is much lower than the one based on charts. In other words, convergence is greater for streaming consumption than it is for pop chart consumption. Consumption therefore did become more similar across countries in the digital era, and greater convergence through streaming further suggests that digitization is promoting convergence. Finally, and perhaps more importantly, consumption is also growing less concentrated by origin of production. This is mainly driven by the decrease in the US world market share and the corresponding increase in other countries' share, as documented above. In other words, it appears that countries' music consumption is becoming more similar with digitization, but their consumption is also becoming more diversified.

2. From Music to Video

Just as in music, digitization has relaxed constraints on movie distribution and, by extension, on movie trade. As a result of both an increase in production and in the emergence of new distribution platforms, consumers worldwide now have access to a much larger set of movies of both domestic and foreign origin.

Netflix is one of the main streaming platforms for video content worldwide. In 2016, they announced their expansion to over 240 countries, allowing their content to be available in most of the world.[29] Like in music, trade in movies also raises important questions regarding the production and distribution of cultural content from smaller countries.[30]

Compared to a music streaming platform like Spotify, Netflix does not offer a comprehensive catalog of movies to its users. While Netflix naturally facilitates

[28] Note that $\sqrt{(0.75-0.1)^2+(0.25-0.9)^2}=0.919$.

[29] As of 2016, Netflix is distributed into 243 sales territories, most of which are countries but some of which are areas within countries. The term "country" is therefore used rather loosely to refer to Netflix distribution territories. See https://media.netflix.com/en/press-releases/netflix-is-now-available-around-the-world and the list of countries at https://help.netflix.com/en/node/14164. Note that Netflix is not available in China.

[30] Many of the questions relevant to the video streaming market are unfortunately hard to tackle given that movie streaming consumption data is essentially not publicly available. Netflix, for instance, is known to be very secretive about their viewership data, so much so that even their content creators do not know how many people watch their shows. See, for instance, http://www.businessinsider.com/netflix-wont-release-streaming-numbers-even-to-creators-2015-11 and https://www.hollywoodreporter.com/news/house-cards-creator-beau-willimon-801280.

access to foreign video content in each country where it is present, the catalog it offers is highly curated and therefore limited. Several reasons explain this fact. First, the rights to distribute existing content are typically country specific, so Netflix cannot offer the same programming in all markets. Second, the model that Netflix follows is highly curated as it entails purchasing content rights for a fixed fee and charging consumers flat fees for unlimited access. Because Netflix incurs costs to include more content but generates revenue only if additional consumers subscribe, they will add content as long as the marginal benefit in subscription revenue cover their marginal costs. Other well-known services follow a similar type of business model, including Hulu, Amazon Prime, and HBO Now. Unlike these curated models, a la carte services offer consumers the option to pay a fee to rent or buy per title, and distributors share revenue with the producers. Amazon Instant Video and Apple iTunes are two of the major a la carte services.[31] Under this business model, distributors have little incentives to limit the amount of content they offer, and catalogs are consequently larger. Table 1 shows how a curated platform like Netflix distributes about 5,000 titles in the US in 2016, while a platform like Amazon Instant offered over 37,000 titles.

Even if Netflix catalogs vary significantly across countries, it is still interesting to assess which origins' repertoires are promoted by the US-based platform. Does Netflix act as a cultural hegemon, mainly using its worldwide presence to distribute US content? Or does it act as a facilitator of free trade, leveling the playing field for smaller markets producers who could not easily distribute their

TABLE 1

WORKS STREAMING IN THE US ON SELECTED PLATFORMS

Platform	*Numbers of movies*	*Numbers of series*	*Total*
Amazom Instant	34,071	3,193	37,264
Netflix	4,186	725	4,911
Apple iTunes	18,657	2,398	21,055
HBO Now	912	73	985
Hulu	3,246	1,537	4,783
Amazon Prime	7,787	487	8,274

Sources: Justwatch.com (retrieved March 2, 2016) and Aguiar and Waldfogel (2017), Table 2.

[31] Note that *Amazon Prime Instant Video* is not the same as *Amazon Instant Video*. The former is an all-you-can-stream service (much like Netflix) available only to Prime members, while the latter is an a la carte service where consumers pay a fee to rent or buy per title. See, for instance, https://www.huffingtonpost.com/2014/02/10/amazon-prime-instant-video_n_4746083.html

content abroad prior to digitization? And how does their curated model, which limits the amount of movies distributed by the platform, affect the way Netflix functions?

Aguiar and Waldfogel (2017) use 2016 data on all country-specific Netflix catalogs and on the origin of the content they carry to explore these questions. The first thing to note is that Netflix makes many of the works from a wide variety of countries available in many other locations. For instance, the Spanish film *The Propaganda Game* is distributed through the platform in 184 countries, the Danish film *Democrats* is distributed in 206 countries, the Hong-Kong film *IP Man 2* is available in 103 countries, and the Thai film *Ong-Bak: The Thai Warrior* is distributed in 63 countries. Basic statistics already give some indication about which origins are promoted on Netflix. Of the 14,390 movies distributed anywhere by the platform in 2016, about 54% were US productions, 9% were from the United Kingdom, 5.9% from France, 3.9% from Canada, and 3.7% from Japan. Spanish movies accounted for 1.3% of the films distributed by the platform. Taken at face value, these figures suggest that Netflix heavily promotes US content. Two things are worth noting though. First, the Netflix catalogs vary substantially across countries. While US consumers had access to about 4,500 movies on the platform in early 2016, Spanish consumers could only access about 1,000 films. Second, and related, not all products are available in all countries. Most movies are available in just 4 countries or fewer.

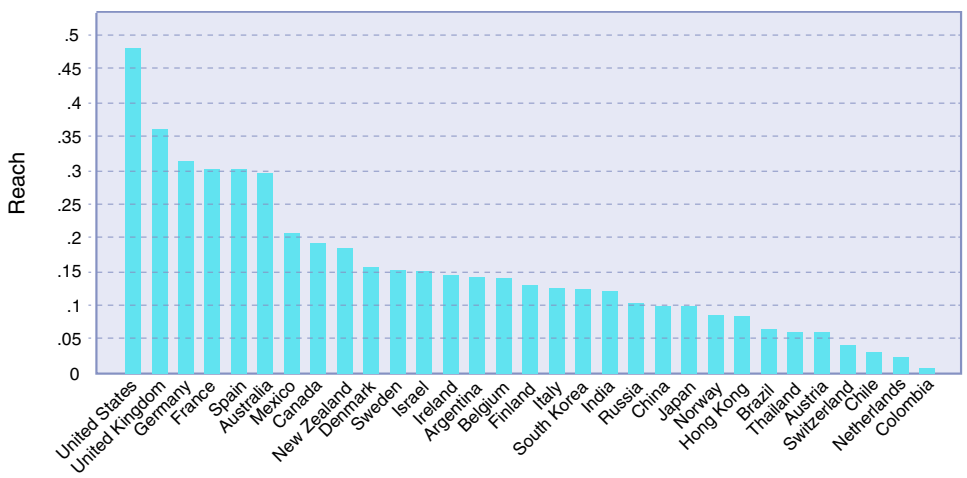
While these descriptive statistics are informative, understanding the extent to which Netflix provides access to an origin's repertoire requires a more refined measure. As a start, one could construct a measure of the coverage provided to a given origin's repertoire by taking into account the total number of existing movies from that given repertoire and the number of countries in which Netflix distributes them. From that perspective, a particular origin's repertoire would have full coverage through Netflix if the platform made all of their movies available in all destinations. But one should also take into account that both countries and movies differ in economic importance. First, not all countries are equal in size. Rather than considering the number of countries in which a movie is made available, one can therefore measure the share of the world population to which the movie is distributed to take country size into account. Second, not all movies from a given repertoire have the same importance. If we had a measure of each movie's value within a given origin's repertoire, we could weight each of the productions by their relative importance. For instance, we could measure the share of the total value of the Spanish repertoire that is made available through Netflix. Finally, characterizing the availability of an origin's repertoire on Netflix naturally requires a point of comparison, and the availability of a repertoire through theatrical distribution is a natural benchmark.

Constructing such measures naturally requires specific data. First, one needs to know the extent of an origin country's repertoire, together with a measure of the importance of each movie included in that repertoire. Aguiar and Waldfogel (2017) gather that information from IMDb.com, an online database related to world films and television programs, among others.[32] Because IMDb provides the number of users rating each movie, the latter can be used a proxy for each work's value and to construct estimates of the share of a given origin country's value included in each distribution channel (Netflix and theater). Second, one needs data on which works are available through Netflix and through theaters. The Netflix data are obtained from unogs.com, which provides 243 country-specific Netflix catalogs together with a link to the corresponding IMDb entry for each title. The theater data come from Box Office Mojo, which provides the list of movies released in theater for 56 countries over the 2008-2014 period. Finally, one can use the population of each country in which each distribution channel operates to measure the share of the world's population which as access to a given movie through each channel.

With all this information, one can construct a measure of the extent to which each distribution channel provides access to an origin's repertoire, which

FIGURE 3

VALUE-WEIGHTED GEOGRAPHICAL REACH, THEATRICAL DISTRIBUTION



Note: Countries with at least 15 movies appearing on Netflix.
Source: Aguiar and Waldfogel (2017), Figure 3.

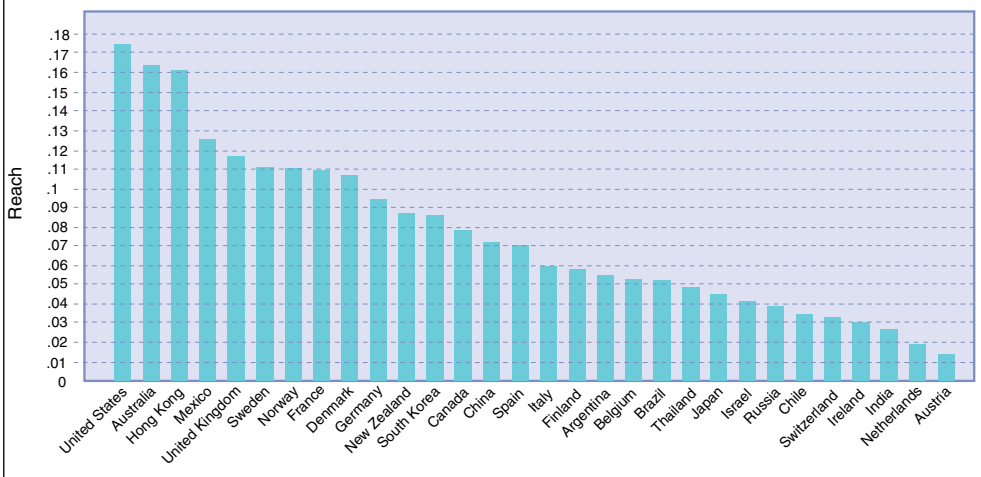
[32] See http://www.imdb.com and https://en.wikipedia.org/wiki/IMDb

Aguiar and Waldfogel (2017) call the value-weighted geographic reach. This essentially provides a measure of the average share of the world's population that movies from a given catalog reaches, and where each movie is weighted by its relative importance within the catalog. For illustration, suppose that a given origin country has 3 movies: one of them is available worldwide on Netflix, the second is only available to half of the world's population through Netflix, and the third is not being distributed anywhere on Netflix. Suppose further that the first movie is twice as important as the two others in the eyes of consumers. In that hypothetical case, the value-weighted geographic reach would be equal to (½×100)+(¼×50)+(¼×0)=62.5 percent.

One can compute that measure for both Netflix and theatrical distribution and for each origin repertoire. Figure 3 reports the reach measure for each repertoire for theatrical distribution in 56 countries and for movies released between 2008 and 2014. The US repertoire is —perhaps unsurprisingly— the one with the largest reach, with a measure of over 0.45. The UK is next at around 0.35, and Germany, France, and Spain follow at about 0.3. Figure 4 reports the reach measure on Netflix for each repertoire using the same countries and underlying population of movies. The US again has the highest reach –about 17.5 percent— followed by Australia (16%), Hong Kong (16%), Mexico (13%),

FIGURE 4

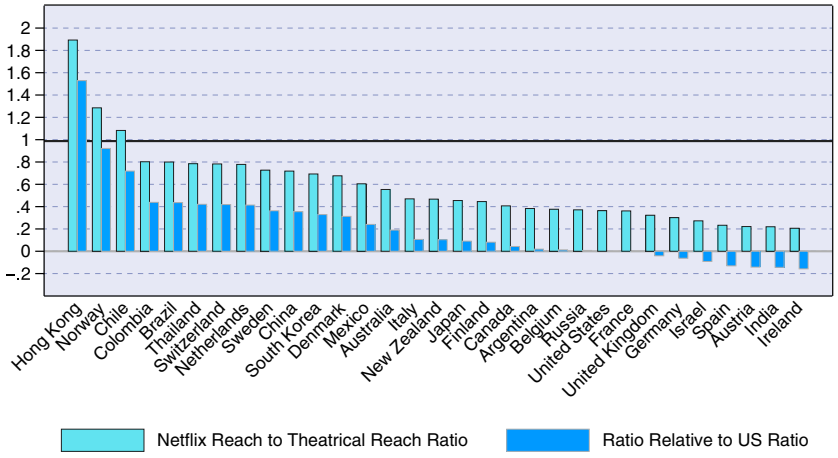
VALUE-WEIGHTED GEOGRAPHICAL REACH, NETFLIX DISTRIBUTION



Note: Countries with at least 15 movies appearing on Netflix.
Source: Aguiar and Waldfogel (2017), Figure 5.

FIGURE 5

RELATIVE VALUE-WEIGHTED GEOGRAPHICAL REACH, NETFLIX VS. THEATRICAL



Note: Countries with at least 15 movies appearing on Netflix.
Source: Aguiar and Waldfogel (2017), Figure 6.

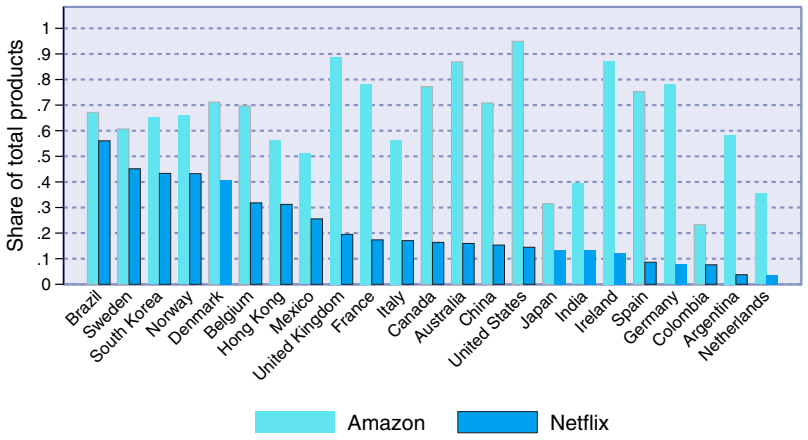
the UK (12%), and Sweden (11%). It is striking that the reach values on Netflix are much lower than on theatrical distribution, indicating that the latter takes a higher share of value to more people for most repertoires. Figure 5 reports the ratio of the Netflix and theatrical value-weighted geographic reach measures. As indicated by the ligth blue bars, the reach of most repertoires is higher through theatrical distribution which is mainly driven by the fact that the highest value movies are largely missing from Netflix. Only Hong-Kong, Norway, and Chile have a larger reach on Netflix than in theater distribution.

The dark blue bars in Figure 5 show the difference between a country's ratio and the ratio of the US repertoire. While most repertoires have lower coverage through Netflix, most countries do relatively better than the US. So even if both theatrical and Netflix distribution of films favor US-origin repertoire, the degree of advantage to US fare is far smaller via Netflix. It is also interesting to see that Netflix seems to be providing an advantage –relative to the US– to smaller-market repertoires. This is reminiscent of the benefits that Spotify seems to be providing to smaller country producers and which we discussed above.

The fact that Netflix carries a relatively small share of the value of an origin's repertoire is likely to be a reflection of its curated business model. Unlike curated services like Netflix, a la carte services tend to offer larger catalogs (see

FIGURE 6

AVAILABILITY OF TITLES ON US AMAZOM AND NETFLIX, BY REPERTOIRE



Note: Countries with at least 15 movies appearing on Netflix.
Source: Aguiar and Waldfogel (2017), Figure 10.

Table 1). It is therefore interesting to see whether these two business models could potentially have an effect on how digitization can facilitate trade. Figure 6 reports the value-weighted share of origin repertoires available in the US at Amazon Instant and at Netflix. Netflix carries an average of around 20% of the repertoires, with higher values reaching 56% for Brazil and 45% for Sweden. Amazon, on the other hand, carries a much larger 65% of the repertoires on average, and over 70% of 10 repertoires. 75 percent of the Spanish repertoire's value is available on Amazon Instant in the US, vs only 9% on Netflix. These figures seem to suggest that the business model employed by *Amazon Instant* can indeed have an effect on how digitization can further help frictionless trade in the movie market.

VI. CONCLUSION

The advent of digitization has initially been seen as a threat to the revenues of many of the content industries. For the recorded music industry, the birth of Napster and the ensuing surge in piracy resulted in a massive decline in recorded music sales, causing major concerns around the potential effects on continued investment in content. While this concern seems warranted, digital technologies have also allowed for an important decrease in the costs of bringing new

products to market. In many of the content industries, the net effect of these two opposing forces has been a dramatic increase in content creation. Because commercial appeal is often hard to predict for cultural products, this increase in production has led to a significant increase in the appeal of newly released products and to substantial welfare benefits. To the extent that quality is also unpredictable for many other products, reductions in the costs of bringing new products to market could have large welfare benefits in other industries as well.

Digitization has also allowed for the emergence of new business models, holding the promise of potentially increasing revenues in many of the content industries. In the recorded music industry, streaming services have been growing in popularity at an astonishing pace and have recently sent global revenues on the path of growth. While US digital downloads revenues have decreased by 21.6% between 2015 and 2016, streaming revenues grew 68.5% and total revenues increased 11.4% to $7.65 billion.

The emergence of streaming platforms in both the music and the movie market have also importantly affected the patterns of trade, raising important questions regarding their effects on content production and consumption patterns. Whereas greater availability of foreign products can potentially allow popular repertoires –such as those of the US– to displace local cultural production in smaller countries, freer trade can also increase the availability of products from countries that have not traditionally produced content with sufficient commercial promise to justify paying the fixed costs of trade. The digital transformation of the cultural industries is still in its infancy, but recent research seems to indicate that digitization has mostly been leveling the playing field, allowing producers from smaller countries reach larger shares of the world market.

The effects of digitization on the content industries remains a fertile ground for future research. Given the large set of products made available to consumers following digitization, one particularly relevant question relates to the ways in which consumers manage to discover the products they find appealing.

BIBLIOGRAPHY

Adams, W. J., and J. L Yellen (1976), "Commodity Bundling and the Burden of Monopoly," *The Quarterly Journal of Economics,* 90: 475-498, http://ideas.repec.org/a/tpr/qjecon/v90y1976i3p475-98.html

Adermon, A., and Ch.-Y. Liang (2014), "Piracy and Music Sales: The Effects of an Anti-Piracy Law," *Journal of Economic Behavior & Organization,* 105: 90—106.

Aguiar, L. (2017), "Let the music play? Free streaming and its effects on digital music consumption," *Information Economics and Policy,* 41*:* 1–14.

Aguiar, L.; Claussen, J., and Ch. Peukert (2018), "Catch Me if You Can: Effectiveness and Consequences of Online Copyright Enforcement," *Information Systems Research.*

Aguiar, L., and J. Waldfogel (2014), "Digitization, Copyright, and the Welfare Effects of Music Trade," *JRC Working Papers on Digital Economy* 2014-05, Directorate Growth & Innovation and JRC-Seville, Joint Research Centre.

— (2016), "Even the losers get lucky sometimes: New products and the evolution of music quality since Napster," *Information Economics and Policy*, 34: 1–15.

— (2017), "Netflix: global hegemon or facilitator of frictionless digital trade?," *Journal of Cultural Economics:* 1–27.

— (2018a), "As streaming reaches flood stage, does it stimulate or depress music sales?," *International Journal of Industrial Organization,* 57: 278-307, http://www.sciencedirect.com/science/article/pii/S0167718717301753

— (2018b), "Quality Predictability and the Welfare Benefits from New Products: Evidence from the Digitization of Recorded Music," *Journal of Political Economy:* 492-524, https://doi.org/10.1086/696229

Anderson, CH. (2006), *The Long Tail: Why the Future of Business Is Selling Less of More*, Hyperion.

Bai, J., and J. Waldfogel (2012), "Movie piracy and sales displacement in two samples of Chinese consumers," *Information Economics and Policy,* 24: 187-196, http://ideas.repec.org/a/eee/iepoli/v24y2012i3p187-196.html

Bakos, Y., and E. Brynjolfsson (1999), "Bundling information goods: pricing, profits and efficiency," *Management Science:* 1613—1630.

Brynjolfsson, E.; Hu, Yu (Jeffrey), and M. D. Smith (2003), "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science,* 49: 1580-1596.

Caves, R. E. (2000), *Creative Industries: Contracts Between Art and Commerce,* Harvard University Press, http://books.google.es/books?id=imfTUHj8uVcC

Danaher, B. (2014), "Testimony of Brett Danaher," http://www.loc.gov/crb/rate/14-CRB-0001-WR/statements/iHeartMedia/Vol%202_02%20

Testimony%20of%20B%20Danaher%20and%20Appendices/2014_10_07_Testimony_of_B._Danaher.pdf, Filed with Copyright Royalty Board, Washington, DC.

Danaher, B., and M. D. Smith (2014), "Gone in 60 Seconds: The Impact of the Megaupload Shutdown on Movie Sales," *International Journal of Industrial Organization,* 33: 1 – 8, http://www.sciencedirect.com/science/article/pii/S0167718713001288

Danaher, B.; Smith, M. D., and R. Telang (2013), "Piracy and Copyright Enforcement Mechanisms," in *Innovation Policy and the Economy*, Volume 14, National Bureau of Economic Research, Inc, NBER Chapters 25-61, http://ideas.repec.org/h/nbr/nberch/12945.html

Danaher, B.; Smith, M. D.; Telang, R., and S. Chen (2014), "The Effect of Graduated Response Anti-Piracy Laws on Music Sales: Evidence from an Event Study in France," *Journal of Industrial Economics,* LXII: 541—553.

Datta, H.; Knox, G., and B. J. Bronnenberg (2017), "Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery," *Marketing Science.*

Goldman, W. (1989), *Adventures in the Screen Trade: A Personal View of Hollywood and Screenwriting,* Grand Central Publishing.

Gomez, E., and B. Martens (2014), Copyright and Geographic Discrimination in the EU Digital Single Market: The Case of Apple iTunes, *Working Paper,* IPTS.

Gourville. J. (2005), The Curse of Innovation: A Theory of Why Innovative New Products Fail in the Marketplace, *Harvard Business School Working Paper,* 16-014.

Handke, Ch. (2012), "Digital copying and the supply of sound recordings," *Information Economics and Policy,* 24: 15-29, http://www.sciencedirect.com/science/article/pii/S0167624512000108

Hui, K-L., and I. Png (2003), "Piracy and the Legitimate Demand for Recorded Music," *The B.E. Journal of Economic Analysis & Policy,* 11, http://ideas.repec.org/a/bpj/bejeap/vcontributions.2y2003i1n11.html

Liebowitz, S. J. (2004), "The elusive symbiosis: The impact of radio on the record industry," *Review of Economic Research on Copyright Issues:* 93—118.

— (2008),"Research Note—Testing File Sharing's Impact on Music Album Sales in Cities," *Management Science,* 54: 852-859, http://ideas.repec.org/a/inm/ormnsc/v54y2008i4p852-859.html

— (2016), "How much of the decline in sound recording sales is due to file-sharing?," *Journal of Cultural Economics,* 40: 13—28, https://doi.org/10.1007/s10824-014-9233-2

McBride, S. (2014), "Written direct testimony of stephan mcbride (On behalf of Pandora Media, Inc)," http://www.loc.gov/crb/rate/14-CRB-0001-WR/statements/Pandora/13_Written_Direct_Testimony_of_Stephan_McBride_with_Figures_and_Tables_and_Appendices_PUBLIC_pdf.pdf, Filed with Copyright Royalty Board, Washington, DC.

Oberholzer-Gee, F., and K. Strumpf (2007), "The Effect of File Sharing on Record Sales: An Empirical Analysis," *Journal of Political Economy,* 115: 1–42.

— (2010), "File Sharing and Copyright," in *Innovation Policy and the Economy,* Volume 10, National Bureau of Economic Research, Inc, http://ideas.repec.org/h/nbr/nberch/11764.html

Peukert, Ch.; Claussen, J., and T. Kretschmer (2017), "Piracy and Box Office Movie Revenues: Evidence from Megaupload," *International Journal of Industrial Organization,* 52: 188—215.

Quan, T. W., and K. R. Williams (2017), Product variety, across-market demand heterogeneity, and the value of online retail, Cowles Foundation *Discussion Paper* nº 2054R.

Reimers, I. (2016), "Can Private Copyright Protection be Effective? Evidence from Book Publishing," *Journal of Law and Economics*, 59*:* 411—440.

Richardson, M., and S. Wilkie (2015), "Faddists, enthusiasts and Canadian divas: broadcasting quotas and the supply response," *Review of International Economics,* 23: 404—424.

Rob, R., and J. Waldfogel (2006), "Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students," *Journal of Law and Economics,* 49: 29-62.

— (2007), "Piracy on the Silver Screen," *Journal of Industrial Economics,* 55(3): 379-395, http://ideas.repec.org/a/bla/jindec/v55y2007i3p379-395.html

Schmalensee, R. (1984), "Gaussian Demand and Commodity Bundling," *The Journal of Business,* 57: S211-30, http://ideas.repec.org/a/ucp/jnlbus/v57y1984i1ps211-30.html

Shiller, B., and J. Waldfogel (2011), "Music for a Song: An Empirical Look at Uniform Pricing and Its Alternatives," *The Journal of Industrial Economics,* 59: 630-660, http://dx.doi.org/10.1111/j.1467-6451.2011.00470.x

Smith, M. D., and R. Telang (2012), Assessing the Academic Literature Regarding the Impact of Media Piracy on Sales, 2012, *Working Paper,* http:dx.doi.org/10.2139/ssrn.2132153

Vogel, H. L. (2014), *Entertainment Industry Economics: A Guide for Financial Analysis,* Cambridge University Press.

Waldfogel, J. (2012), "Copyright Protection, Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster," *Journal of Law and Economics,* 55: 715-740.

— (2013), *Digitization and the Quality of New Media Products: The Case of Music in Economics of Digitization,* University of Chicago Press, http://www.nber.org/chapters/c12996

— (2015), *Digitization and the Quality of New Media Products: The Case of Music,* University of Chicago Press: 407-442, http://www.nber.org/chapters/c12996

— (2016), "Cinematic Explosion: New Products, Unpredictabilty and Realized Quality in the Digital Era," *The Journal of Industrial Economics,* 64: 755—772, http://dx.doi.org/10.1111/joie.12117

— (2017), "How Digitization Has Created a Golden Age of Music, Movies, Books, and Television," *Journal of Economic Perspectives,* 31: 195—214.

Waldfogel, J., and I. Reimers (2015) "Storming the gatekeepers: Digital disintermediation in the market for books," *Information economics and policy,* 31: 47—58.

Waldfogel, J.; Aguiar, L., and E. Gómez (2017), "Does Digitization Threaten ocal Culture? Music in the Transition from iTunes to Spotify," *Working Paper.*

Wlömert, N., and D. Papies (2015), "On-demand streaming services and music industry revenues – Insights from Spotify's market entry," *International Journal of Research in Marketing,* http://www.sciencedirect.com/science/article/pii/S0167811615001226

Zentner, A. (2006), "Measuring the Effect of File Sharing on Music Purchases," *Journal of Law and Economics,* 49(1): 63-90, 2006.

THE ECONOMICS OF THE GIG ECONOMY – WITH AN APPLICATION TO THE SPANISH TAXI INDUSTRY

Mateo SILOS RIBAS[1]

Abstract

The gig economy is one of the latest byproducts of modern technological progress. It is a process of disruptive innovation that is transforming markets and society as a whole. This article reviews the economics of the gig economy and explores its main economic and regulatory implications. Throughout the article, the Spanish taxi industry is used as an example of how traditional regulations are hampering the development of the gig economy and imposing substantial costs on consumers and society.

Key words: Technological progress, peer-to-peer markets, regulation, competition, taxi.

JEL classification: D40, L11, L43, L86.

[1] The views expressed in this article are the sole responsibility of its author and do not necessarily represent those of his employer.

I. INTRODUCTION

In the last decade, peer-to-peer markets have flourished in many sectors of the economy, ranging from rental accommodation to transport services, professional services, financial services or household tasks, among others. These markets help buyers and sellers find each other and allow them to share assets or resources through the Internet. They are often referred to as the "gig", "sharing" or "collaborative" economy and they constitute one of the latest byproducts of modern technological progress.

This article reviews the economics of the gig economy and explores its main economic and regulatory implications. Throughout the article, the Spanish taxi industry is used as an example of how traditional regulations are hampering the development of the gig economy and imposing substantial costs on consumers and society.

Peer-to-peer markets are economic platforms that internalize transactions costs and connect different groups of peers willing to undertake economic transactions. They use a myriad of modern technological innovations to build virtual online marketplaces that enable individuals (peers) to share assets in a relatively easy, efficient and reliable way. Transactions between peers are relatively short term, rely on spot markets and often unlock underutilized assets. Assets are exchanged when their owner is not using them for personal consumption.

Since its early stages, the gig economy has experienced remarkable growth and has become very popular. Throughout the world, consumers are increasingly using providers such as Airbnb, Uber or TaskRabbit, to name a few. These are very novel companies that did not exist a decade ago and are now increasingly present in our lives.

Sharing underutilized assets by individuals was certainly possible before the emergence of modern peer-to-peer markets. However, exchanges were relatively scarce. High transaction costs and informational problems usually led to risky and costly transactions, making them very infrequent and often restricted to exchanges with relatives or friends. Modern peer-to-peer markets overcome these problems. This is why they have become so popular and successful. Peer-to-peer markets lower transaction costs, improving the matching of supply and demand. They do this very effectively. In addition, they ensure trustworthy transactions. Trust is a crucial factor to create the network effect between buyers and sellers. These markets build trust by solving informational problems through identity verification systems, feedback and evaluation tools and reputational mechanisms. It is surprising how effective these mechanisms are in enabling asset sharing transactions between strangers. As transactions have become

cheaper, faster and safer than in the past, they have flourished and become pervasive.

The gig economy affects how services are produced, distributed and consumed. This should come as no surprise, as it has happened in the past with other technological innovations. Peer-to-peer markets lower entry barriers, enhance competition and disrupt markets. Its emergence has already transformed many industries forever. The case of taxi services is a paramount example. For decades, this industry has been regulated as a legal monopoly across countries. Quantity restrictions and regulated fares have shielded incumbent players from competition, ensuring them a peaceful and "quiet life" (Hicks, 1935). In terms of innovation, the taxi industry has remained relatively unchanged during the last decades until very recently, when technological change has shaken the industry. In the last ten years, ride-sharing platforms like Lyft or Uber have succeeded in using new technologies to match drivers and passengers in effective and innovative ways, increasing efficiency, expanding supply and fostering competition in terms of price, quality and variety. They have emerged as a substitute for traditional providers, offering consumers a new experience in urban transportation services. Where these platforms are providing service with an effective ability to compete, they have intensified competition and massively attracted drivers and passengers. In certain cases, they have fully eroded the monopoly position of incumbent operators. Markets for rental accommodation, financial services, professional services, household tasks, touristic guides or even babysitting are also being disrupted by the gig economy.

The potential benefits of the gig economy for society are substantial. Improvements in efficiency and increased competition lead to lower prices, higher quality, more variety, increased innovation, higher productivity and, ultimately, more economic growth and higher living standards. In sum, better lives for all. Unfortunately, existing regulations are hampering the development of peer-to-peer markets across sectors. Again, the taxi industry is an interesting case. Services provided by a company like Uber are facing severe restrictions, or even straight bans, across cities and countries throughout the world. Spain constitutes a formidable example. Incumbent players in the taxi industry are reacting strongly against innovative entrants and lobbying to maintain or reinforce existing regulations so as to stifle healthy competition and protect the monopoly regime in taxi services.

When existing regulations were first enacted, regulators often advocated them on welfare grounds. Their intended purpose, so they said, was to correct market failures and enhance welfare in comparison to the outcome which would have prevailed in an unregulated market. Although many sectoral regulatory provisions were just a byproduct of anticompetitive rent seeking –quantity

restrictions in the taxi industry are a classic example– the fact is that peer-to-peer markets solve or mitigate through innovative ways the very same market failures that allegedly motivated existing regulations in the first place. Their success and popularity are precisely rooted in their capacity to make markets work in an efficient fashion, by lowering search costs and solving informational problems. As peer-to-peer markets contribute to solving or mitigating market failures, many regulatory provisions intending to correct them become unjustified on welfare grounds and should be removed or modified. The gig economy gives society the opportunity to improve how it regulates markets and improve welfare for all.

Regardless of the potential benefits of the gig economy, reforming existing regulations is proving very hard. Regulators are behaving as they usually do when innovation thrives and disrupts markets. First, they are neglecting the benefits of entry and competition. Second, they are worried about the adverse effects of entry on incumbent players. Third, they are working to preserve the regulatory *statu quo,* opposing beneficial reform for society. Sadly, the sharing economy is yet another example of how vested interests, lobbying efforts and regulatory capture can block the path to pro-competitive reform, depriving society of higher levels of prosperity brought by modern technological progress.

The article is organized as follows. Section II provides an overview of the basic economics of the gig economy. Section III reviews the empirical literature on the effects of the gig economy on markets, competition and welfare. Section IV explores how traditional regulatory frameworks may hinder the development of gig economy business models, focusing on the case of taxi services in Spain. Section V addresses the costs that some traditional regulatory frameworks impose on consumers and society as a whole, using the consumer welfare loss arising from taxi regulations in Spain as an illustration. Section VI concludes.

II. SOME BASIC ECONOMICS OF THE GIG ECONOMY

In the last decade, peer-to-peer markets have flourished in many sectors of the economy, such as rental accommodation, transport services, professional services, financial services or household tasks. These markets help buyers and sellers find each other and allow them to share assets or resources through the Internet. They are often referred to as the "gig", "sharing" or "collaborative" economy and they constitute one of the latest byproducts of modern technological progress.[2]

[2] Throughout the article the terms "peer-to-peer markets", "gig economy", "sharing economy" and "collaborative economy" are used interchangeably.

Although peer-to-peer markets are diverse, they also have some common features (see for instance Fraiberger and Sundararajan, 2017; Einav, Farronato and Levin, 2016; Spence, 2015; Sundararajan, 2014a; Sundararajan, 2014b; The Economist, 2013). First, they are economic platforms that internalize transaction costs and connect different groups of peers willing to undertake economic transactions. Peers are individual buyers and sellers who can be on either side of the market, a reason why they are sometimes called *prosumers*. Second, these markets are technology driven. They take advantage of the Internet, related information and communication technologies, market design mechanisms and big data to build online marketplaces where individuals meet and transact. Third, they enable individuals to share assets in a relatively easy, efficient and reliable way. Thanks to technological progress, assets such as homes, boats, sewer machines or lawnmowers become disaggregated and consumed as services, on a massive scale. Fourth, transactions undertaken in peer-to-peer markets often unlock underutilized assets. Assets are exchanged when their owner is not using them for personal consumption. Fifth, transactions are short term and rely on spot markets. Sixth, peer-to-peer markets lower entry barriers for individual sellers and allow them to compete against incumbent players across markets in the economy.

A good example of a peer-to-peer market is UberX, one of the first services launched by Uber[3] in the United States and a true symbol of the gig economy. UberX is a service option allowing individuals to drive for Uber using their own car, fulfilling a set of background checks and car requirements. In the beginning of 2013, this service was present in more than 35 cities,[4] often facing several regulatory problems.

UberX is an iconic example of a peer-to-peer market. First, it is a platform-based application that connects drivers and passengers seeking rides in an

[3] Uber is a ride-sharing platform that uses modern technology to connect drivers and users of urban transportation services. Uber operates in more than 600 cities worldwide (Wikipedia). It was officially launched in San Francisco in 2011. Passengers pay a fare based on the distance of their trip and the time taken to complete the trip while drivers receive this fare minus a service fee paid to Uber. In many cities, Uber is well-known for using a dynamic pricing mechanism: Uber adjusts its prices using a real time dynamic algorithm known as *surge pricing*. When demand increases relative to supply in a given area, the price changes to match supply and demand. Depending on the city, Uber offers many types of services, such as UberPool –a low cost alternative, that allows individual passengers to share a ride with other passengers– UberX –a peer-to-peer market where individual drivers share their own cars, also known as UberPop in some European cities– and UberBlack, an option offering a ride experience with relatively high-quality cars and licensed drivers. Uber uses two-way evaluation systems through which drivers rate and evaluate passengers and vice-versa and also other tools –such as external regulations on drivers– to ensure the quality of its services. Uber faces competition from many similar applications. The most well-known is Lyft, mainly operating in the United States and more recently in Canada. Uber and Lyft compete both for drivers and passengers in many cities.

[4] Source: Wikipedia.

urban area. Second, it is essentially based on new technologies. The matching between drivers and passengers is undertaken through an algorithm. Internet, Global Positioning System (GPS) technologies, smartphones, market design mechanisms and big data lie at the heart of UberX's business model. In fact, UberX would not exist without all these technological innovations, especially the smartphone. Third, UberX allows individuals to share their own asset –a car– when they are not using it for personal consumption so that other individuals can consume it as a service. Transactions are undertaken in a relatively easy, efficient, convenient and reliable way.[5] Fourth, UberX unlocks a marvelous example of dormant physical capital: cars. Most of their life, cars remain unused.[6] Fifth, rides in UberX are agreed on spot transactions that are short-term in nature. In 2015, the average Uber trip distance in the United States was 6.4 miles.[7]

Since its early stages, the gig economy has experienced a remarkable growth and has become very popular. Since the launch of Airbnb in 2008, there have been over 200 million total guest arrivals around the world.[8] On May 20 2017 Uber hit the cumulative figure of 5 billion rides since its launch.[9] In Europe, five key sectors of the gig economy generated revenues of approximately 4 billion euros and facilitated a transaction value of 28 billion euros in 2015.[10] Between 2013 and 2015, revenues generated by the gig economy in Europe tripled (PwC, 2016). Throughout the world and across sectors consumers are increasingly using providers such as Airbnb, Uber or TaskRabbit, to name a few. These are very novel companies which did not exist a decade ago and today are becoming increasingly present in our lives. In 2016, around a third of European consumers had heard about the gig economy and 5% of them had effectively participated in gig economy platforms.

Sharing underutilized assets was certainly possible before the development of modern peer-to-peer markets (Horton and Zeckhauser, 2016). Renting by

[5] Since UberX was launched, consumers and drivers have increasingly used the platform. The number of UberX drivers has experienced an exponential growth in the United States since 2012 (Hall and Krueger, 2017). Uber is competing intensively in some cities and displacing taxi services. An interesting indicator is how peer-to-peer platforms' entry is impacting on the value of taxi licenses –medallions– in many cities, such as Chicago and New York City (Bacgchi, 2017) or Sidney and Melbourne (OECD, 2017), among others.

[6] In 2009, the average minutes that drivers spent driving a private vehicle in a typical day in the United States was 56 minutes, less than an hour (US Department of Transportation, 2011). Indeed, the insight behind many business models in the gig economy –not just UberX– is that the world is full of underutilized assets such as cars (Spence, 2015). Technology has finally enabled society to unlock all this dormant physical capital (Sundararajan, 2014b).

[7] Data from the last quarter of 2015. Source: SherpaShare, 2016.

[8] Source: Airbnb.

[9] Source: Uber.

[10] PwC (2016) focuses on the following five key sectors: peer-to-peer accommodation; peer-to-peer transportation; on-demand household services; on demand professional services; collaborative finance.

consumer owners has existed for centuries. However, it was generally restricted to expensive goods –such as big mansions or yachts– within a long-term lease. Moreover, most of these trades took place between relatives or friends, usually without payment. New rental markets in the gig economy are wide range, involve massive transactions between strangers and are often subject to payments.

Although the economic problem solved by peer-to-peer markets is not new, exchanges of underutilized assets between individuals have traditionally been very scarce and infrequent due to several reasons (Horton and Zeckhauser, 2016). First, the existence of search costs, that buyers incur in to find sellers. Second, informational problems, that both buyers and sellers confront. Agents may lack information regarding the assets to be shared or the prospective buyer. Information may be imperfect or asymmetric. Third, until very recently, individuals lacked firm-like resources, such as marketing budgets, business expertise, ways of dealing with payments, contract procedures, insurance policy or a brand.

Peer-to-peer markets take advantage of modern technology to lower search and information costs and overcome the problems that traditionally hampered asset sharing transactions between individuals (Horton and Zeckhauser, 2016). In addition, they rely on the scientific progress made in the area of market design mechanisms (see Vulkan, Roth and Neeman, 2013), for a comprehensive review), as well as in the stock of knowledge and experience accumulated after two decades of e-commerce and online marketplaces. Indeed, the first generation of online marketplaces brought by the Internet, such as Amazon or eBay, had to overcome informational problems inherent to online transactions which are similar to those that modern peer-to-peer markets have to deal with. Lastly, peer-to-peer markets provide individuals with firm-like resources. Compared to individuals, platforms generate economies of scale in producing all these tasks, which would be too costly for an individual producer. For instance, platforms enable individuals with spare rooms in their houses to become entrepreneurs in the lodging industry. A platform like Airbnb allows its hosts to contact millions of potential customers, communicate with them, sign a contract, and manage payments. Individuals from Tokyo or Anchorage rent rooms to strangers in La Coruña or Ankara, through safe, fast and reliable transactions. Today, these types of transactions amount to millions. However, they were very infrequent, almost non-existent, two decades ago.

Peer to peer markets have features in common with the first generation of online marketplaces, such as Amazon or eBay. First, they take advantage of new technologies, such as the Internet, modern electronic payment systems,

GPS, smartphones or online reviews. Second, they must build trust, as trust constitutes as a key factor to succeed in digital markets. Ensuring trust is what makes possible the sale and resale transactions between strangers that take place in a platform like eBay over a wide range of goods, including cars, planes or even medical equipment. eBay and other online marketplaces were pioneers in using evaluation and feedback systems to solve informational problems. Indeed, this is one of the reasons explaining their success. Third, peer-to-peer markets make markets thick and global (Einav, 2015). A market can be thin in several dimensions, such as product definition or geography. For instance, Uber has made urban transportation markets thick and global, moving them away from their thin and local structure. In physical markets, market thickening is undertaken through geographical or temporal coordination mechanisms (Horton and Zeckhauser, 2016). As online marketplaces generally lack these, they resort to creating taxonomies, classifying goods, making use of search algorithms and user recommendation systems. Fourth, peer-to-peer markets are multisided markets (Belleflame, 2017). Multisided platforms connect different groups of users willing to find each other and interact. One of their distinct features is the existence of indirect network externalities between groups: the value that users in one group obtain from the platform increases with the number of users in the other group(s).

Most of the markets with indirect network externalities are multisided markets (Rochet and Tirole, 2003 and 2006). A market with indirect network externalities is a multisided market if a platform can cross-subsidize between groups. The platform's aggregate demand and profits depend on the price structure, that is, on how the price is distributed on each side. The price structure is non-neutral as regards to the network effect. Whether a platform becomes successful depends to a great extent on whether it has achieved an effective price structure to get both sides on board (Caillaud y Jullien, 2003). In most cases, the price structure is asymmetric, and one side pays less, or even zero. Multisided markets constitute a very broad category of markets, encompassing both offline platforms –such as downtown Manhattan, credit cards, airports or newspapers and radios– and online platforms –such as Google, eBay, Uber or Airbnb. Peer-to-peer markets are a subcategory of multisided market. They connect producers and consumers, such as hosts and guests in the case of Airbnb or drivers and passengers in the case of Uber.

Despite some common features, peer-to-peer markets are distinct from the first generation of online marketplaces. First, transactions are focused on short term rental or service provision, rather than in the sale or resale of goods with a transfer of ownership between parties (Fraiberger and Sundararajan, 2017). Second, trust between parties is relatively more important given the nature of the transactions (Einav, 2014). Transactions are more intimate since

they relate to individuals sharing their homes or cars with strangers. As a result, feedback and reputational mechanisms become much more sophisticated and effective. Third, on the technological side, peer-to-peer markets not only rely on the Internet but also on technologies such as GPS, smartphones or social networks (Einav, 2014). Fourth, individuals and in particular *prosumers* are more prevalent in today's peer-to-peer markets than in the first generation of online marketplaces. Firms may also be present in either the demand or the supply side of the gig economy, but to a lesser degree than in the first generation of online marketplaces.

Einav, Farronato and Levin (2016) explore the economic properties of peer-to-peer markets and the strategies they need to implement to become successful. The main goal of peer-to-peer markets is to create trade between a large number of dispersed and fragmented buyers and sellers. To achieve this goal, they need to develop a market design strategy that addresses a series of challenges in terms of matching, pricing and trust.

Peer-to-peer markets help buyers and sellers find each other, tackling a search problem. In addition, they must set up a price mechanism to balance supply and demand in the virtual marketplace. In addressing matching and pricing problems, peer-to-peer markets face a trade-off. On the one hand, information is dispersed and preferences are heterogenous. Therefore, peer-to-peer markets have to elicit and incorporate dispersed information, facilitating search. On the other hand, for the market to be successful, peer-to-peer markets must ensure that user experience is convenient. Hence, they must keep transaction costs low. The existence of this trade-off gives rise to a plurality of market settings in the gig economy.

In the case of rental accommodation, heterogeneity is extremely important. Consumer demand is highly heterogenous. Hence, it would be impossible to establish a common ordering regarding accommodations across individuals. Supply also exhibits a high degree of heterogeneity both in characteristics and costs. In this context, a decentralized market setting emerges as the best way to match supply and demand. Buyers choose from a diverse range of sellers. For instance, Airbnb allows users to specify their preferences and presents them with results, but ultimately allows users to choose their preferred provider. As both matching and prices relate to an information problem, prices are also set in a decentralized fashion. Suppliers post their prices and consumers choose. Prices are dynamic and contingent on demand and supply conditions, a feature often observed in peer-to-peer markets.

In the case of urban transportation, the market design is substantially different. Consumers value an immediate dispatch. Therefore, supply and

demand must match in real time and as fast as possible. Users value both the safety and promptness of service and do not have a strong preference for other characteristics. Demand and supply heterogeneity is less relevant. Therefore, a centralized system emerges as the most efficient way to address both matching and pricing problems. UberX is a good example. Users ask for rides, but do not choose a particular driver. Drivers accept rides, but do not choose passengers. The system dispatches drivers to passengers in a centralized fashion, keeping transactions costs low and user experience convenient. Prices are also contingent on demand and supply conditions but set in a centralized fashion. There is a fee per ride, reflecting distance and time.

The third challenge that peer-to-peer markets have to address is ensuring trustworthy transactions (Einav *et al.,* 2016; Horton and Zeckhauser, 2016; Spence, 2015; Sundararajan, 2014a). Trust is a relatively acute problem in the gig economy because consumers have the opportunity to misuse valuable physical capital, such as homes or cars, and transactions are more intimate. Trust is what makes possible that asset sharing transactions between strangers –individuals who do not know each other and have never met– take place on a massive scale.

There are several ways to build trust (Einav *et al.,* 2016): up-front inspection, external enforcement and reputation. Peer-to-peer markets resort to all of them. For instance, they may impose external regulations. UberX drivers must comply with minimum quality standards determined by the platform. Airbnb offers hosts the possibility of certifying photos of their properties and verifying the identity of hosts and guests. However, reputational mechanisms constitute by far the major tool of peer-to-peer markets to ensure trustworthy transactions. In fact, the relevance of these mechanisms in peer-to-peer markets is one of their distinct features in comparison to traditional physical markets. Reputation systems can be considerably sophisticated. For instance, Uber uses a two-way evaluation and review system where passengers evaluate drivers and vice versa. Drivers and passengers can be expelled from the platform if their ratings fall below a minimum threshold.

Einav *et al.* (2016) build a theoretical model to analyze the impact of peer-to-peer markets on traditional industries, such as the lodging or the urban transportation industry. They model a market where peer producers compete with traditional operators who make up-front investments in dedicated capacity. Flexible suppliers do not incur in up-front fixed costs. In a market with dedicated capacity only, capacity is fixed, so there is more price variability. If there is also peer production, capacity becomes more flexible and supply elasticity increases. As supply becomes more responsive to demand variations, prices fluctuate less. The empirical literature confirms that peer-to-peer markets increase supply

elasticity. For instance, Brodeur and Nield (2016) explore whether Uber's entry in New York City has made it easier to get a taxi in rainy days. They find that when it rains the number of Uber rides increases by 25% while the number of taxi rides increases by 4% only. Zervas, Proserpio and Byers (2017) provide evidence of how Airbnb increases supply elasticity in the lodging industry.

Einav *et al.* (2016) explore the conditions that make peer production more likely. First, relative costs between dedicated supply and flexible (peer) supply. If up-front investment is relatively low, peer-to-peer production will be less likely. Second, visibility and advertising costs. As these costs increase, the likelihood of peer-to-peer entry will be lower. Third, demand variability. In markets where demand is variable, the efficient form of production entails having some capacity that only operates at peak times. Having dedicated capacity that only operates at peak times is more expensive than having flexible suppliers providing short-term supply at peak times. As demand variability is high both in the lodging and the urban transportation industries, there is an economic rationale for flexible supply to have a role in these markets. The empirical evidence confirms that when technology has made it possible, flexible producers have populated these industries thanks to platforms like Airbnb in the lodging industry or Lyft or Uber in the urban transportation industry. Indeed, these industries are the ones that have been more affected and disrupted by the gig economy up to date.

Sundararajan (2014a) highlights some of the economic effects of peer-to-peer markets. First, they expand consumption, creating new experiences for consumers and increasing variety. For instance, thanks to Airbnb, consumers benefit from a more differentiated supply and a wider range of options, which were not available before. Second, they increase productivity, enabling a more efficient use of physical or human capital. For instance, UberX increases the productivity of cars, unlocking them from their dormant state. Third, they foster entrepreneurship and innovation. As they lower entry barriers and the risks of becoming an entrepreneur, they allow individuals to incur in projects that otherwise they would not have pursued. Fourth, they cause shifts in asset markets because they affect asset liquidity and consumption and investment patterns.

Peer-to-peer markets facilitate market entry through different channels, most of them already discussed. First, they lower entry barriers and allow peer producers to compete against incumbent operators. Peer-to-peer platforms allow peers to share fixed costs –such as advertising costs– lowering the efficient firm size. Reputation systems significantly contribute to easing entry. For instance, cab drivers in London have traditionally spent years studying in order to obtain a black cab license. However, the application process to become an Uber driver takes hours or days (Einav *et al.,* 2016). Second, they reduce transaction costs. Thanks to new technologies and modern payment systems, it

is much easier for buyers and sellers to undertake economic transactions. Third, they reduce search costs, for instance, the cost of finding a driver at night or an available apartment in a given date. Fourth, they allow individuals to have firm-like resources (Horton y Zeckhauser, 2016). Fifth, they solve informational problems. In sum, peer-to-peer markets make it easier to become a seller. As a result, they have the potential to substantially expand supply.

As they ease entry and expand supply, they increase competition (Einav *et al.,* 2016; Zervas, Proserpio and Byers, 2017; Stallibrass and Fingleton, 2016; CNMC, 2016b;[11] FTC, 2013). Peer-to-peer markets emerge as an alternative and innovative provider in the market. Platforms such as Airbnb or Uber foster competition and benefit consumers in several ways, such as lower prices, more quantity, increased variety or more innovation. Increased entry and competition of new business models triggers beneficial competitive reactions by incumbent players. Incumbents may respond lowering their prices, improving their productive efficiency or innovating more by offering new products and services. The next section reviews the empirical evidence on the effects of peer-to-peer markets on markets and welfare.

III. IMPACT OF THE GIG ECONOMY ON MARKETS AND WELFARE: THE EMPIRICAL EVIDENCE

This section reviews the most relevant empirical literature assessing the impact of peer-to-peer markets on markets and welfare. The literature ranges from relatively descriptive analyses to more sophisticated research focusing on the causal impact of peer-to-peer markets on markets and welfare. Overall, the empirical literature finds a positive impact of peer-to-peer markets. They improve efficiency, lower entry barriers, foster competition, reduce prices and enhance welfare. The review provided below covers studies both in the lodging and the urban transportation industries, although a special focus is given to those assessing the impact of peer-to-peer markets on the latter.

Cramer and Krueger (2016) analyze the efficiency of peer-to-peer markets by comparing capacity utilization rates between UberX drivers and taxi drivers

[11] CNMC (2016b) is a market study on the sharing economy elaborated by the Spanish Competition and Markets Commission (CNMC). This study was a pioneer initiative in the competition policy field. Its scope of analysis encompasses the lodging industry, interurban bus services and the taxi industry. The study contributed to fostering the debate on the sharing economy both at international and national level in Spain. Its policy recommendations constitute a formidable package that would allow for an efficient entry of gig economy business models in the Spanish economy. Maudes, Sobrino and Hinojo (2017) provide an overview of its intellectual approach in terms of economic analysis and policy principles.

in the United States. They measure capacity utilization in two ways: (i) the percentage of time that drivers have a customer in their car or (ii) the percentage of miles driven with a customer in their car. They provide estimates of capacity utilization rates for a group of cities in the United States[12] using different data sources.

Their results suggest that capacity utilization is higher in UberX than in the taxi segment. First, on average, UberX drivers have a passenger in their car half of the time. This result is similar across cities. On the contrary, taxi drivers have a passenger in the car between 30% and 50%, depending on the city. Second, UberX drivers are more productive in terms of miles driven with a passenger in their car. For instance, in the cities of Los Angeles and Seattle, UberX drivers have a passenger for 64.2% and 55.2% of their miles driven, whereas the utilization rates for taxis in those cities stand at 40.7% and 39.1%. Using the average across cities and measures, capacity utilization is 38% higher for UberX drivers. Ignoring fixed costs and assuming linear fares, these results mean that on average UberX drivers could establish a price 28% lower than taxi fares and earn the same revenue per hour. In the city of Los Angeles, where the differential in terms of capacity utilization is higher, fares could decrease by 37%.

The authors suggest different factors that may explain UberX's higher capacity utilization rates. First, UberX uses better technologies to match drivers and passengers. Taxi operators still rely on dispatch technologies dating back to the 1940s or, alternatively, on cruising and direct street hailing. Second, UberX has achieved a larger scale, creating network efficiencies. Third, inefficient taxi regulations. For instance, those preventing taxi drivers from dropping off a customer in a location outside their jurisdiction and picking up a passenger from that location in the returning trip. They must return empty. Fourth, UberX uses innovative pricing mechanisms –namely, surge pricing– that increase supply's responsiveness to market conditions. In comparison to regulated taxi fares, these pricing mechanisms are more efficient and improve how supply and demand match.

Peer to peer markets disrupt industries and intensify competition. Zervas, Proserpio and Byers (2017) estimate the effect of Airbnb in the traditional hotel industry. In particular, they assess the impact of Airbnb's entry in the Texas hotel market –namely on hotel room revenue– and provide evidence on substitution patterns and competitive effects. The article uses data collected by authors from Airbnb and data on hotel room revenue from administrative public records of approximately 3,000 hotels in Texas. To estimate the causal impact of Airbnb's

[12] Boston, Los Angeles, New York, San Francisco and Seattle.

entry they use a differences-in-differences strategy, using Airbnb's entry as the intervention (treatment) variable. As they have data for several years and Airbnb's entry in Texas was sequential –both in time and space– they can exploit this spatiotemporal variability to estimate the causal effect of Airbnb's entry on hotel room revenue.

They obtain several empirical results. First, Airbnb's entry has a negative impact on hotel room revenue. In the authors' preferred econometric specification, each additional 10% increase in the size of the Airbnb market leads to a 0.39% reduction in hotel room revenue. In the case of Austin, where Airbnb supply is relatively higher, the causal impact on hotel revenue is in the 8%–10% interval. Second, Airbnb's entry triggers competitive reactions by incumbent hotel operators. In the short run, the authors find a small reduction in hotel occupancy rates and a significant decrease in hotel prices. Third, Airbnb's entry makes supply more elastic. Airbnb flexibly scales instantaneous supply in periods of high demand, curtailing hotels pricing power in peak periods. Fourth, Airbnb has a differentiated effect across hotel categories. Low-price hotels and independent hotels are relatively more affected by Airbnb's entry. On the contrary, relatively expensive hotels, chain hotels and hotels offering congress amenities are less affected.

As regards to peer-to-peer markets in urban transportation, Canada Competition Bureau (2015) provides descriptive evidence on the short-term market dynamics after the entry of platforms like Uber and Lyft. The analysis highlights the positive impact of peer-to-peer markets on variables that affect consumer welfare. First, they reduce waiting times, because they increase aggregate vehicle supply, use new technologies that improve the matching of supply and demand and enhance the quality and speediness of drivers' performance through feedback and reputation systems. For instance, in the Canadian cities of Toronto and Ottawa waiting times are lower in Uber rides (2-4 minutes in Toronto; 3.7 minutes in Ottawa) than in taxi rides (9; 5-15). Second, they lower prices. Platforms are cheaper than traditional taxi operators[13] –this is one of the main reasons why consumers substitute them for taxi services– and they trigger positive competition reactions by incumbent taxi operators. After Uber's entry in Toronto, the city council lowered the regulated base taxi fare by approximately 25% to enable taxi operators to better compete against Uber. Third, they improve consumer experience and convenience. For instance, customer surveys in Ottawa indicate that overall customer service experience in Uber is higher than in traditional taxis (City of Ottawa, 2015). Competitive reactions by incumbents are also relevant in the case of non-price variables, such as customer experience and convenience. In 2015 in London, a month after

[13] See also Silverstein (2014) for several cities in the United States.

the High Court ruled that Uber was allowed to provide urban transportation services, London's black taxis started to accept card and contactless payments.[14]

As regards to service availability, Bialik, Flowers, Fischer-Baum and Mehta (2015) provide evidence on how peer-to-peer markets in urban transportation services improve supply's response to demand segments which have traditionally been relatively underserved by the taxi monopoly, such as urban peripheral areas. Using data for New York City they find that traditional taxis tend to concentrate relatively more in central locations, leaving relatively under-served outer urban areas. In 2014, 22% of Uber rides started outside Manhattan, compared to a figure of 14% in the case of traditional yellow and green taxis. This evidence suggests that entry of new platforms increases the availability of supply in relatively more distant areas. Increased and more efficient urban connectivity fosters economic and labor opportunities for individuals living in those areas, enhances agglomeration economies and fosters economic growth.

As peer-to-peer markets intensify competition, they contribute to dissipating inefficient rents. An interesting example is their impact on the value of taxi licenses (medallions). The value of taxi licenses reflects the discounted value of the future revenue stream of supra-competitive rents in a monopoly environment. In an open competitive market, a taxi medallion would have no value. Hence, if entry and price competition increase, the value of medallions will decrease, as monopolistic rents will dissipate. This is what has happened in many taxi markets where peer-to-peer markets have entered with an effective ability to compete. Bagchi (2017) analyzes the changes in taxi medallion prices in New York City (period 2009-2016) and Chicago (2007-2016). The paper finds a drop of roughly 50% in medallion prices in New York City and roughly 80% in Chicago from their peak in 2013-2014 to 2016. In addition, the analysis suggests a positive correlation between medallion prices and taxi revenues and a negative correlation between medallion prices and proxies for the intensity of adoption of Uber and Lyft in both New York City and Chicago.

Finally, some studies have focused on quantifying the value of peer-to-peer markets for consumers. One interesting study is Cohen, Hahn, Hall, Levitt and Metcalfe (2016). They estimate the consumer surplus generated by UberX in the United States. They provide estimates of consumer surplus in four United States' cities and "back-of-the envelope" calculations for the United States as a whole.

[14] Source: "London's black cabs to accept contactless payments as fight against Uber intensifies", *The Telegraph,* November 2015.

They use individual data from UberX rides. They rely on Uber's surge price algorithm to identify changes in prices.[15] Through surge pricing, UberX uses local demand and supply conditions to produce equilibrium prices. Changes in market conditions, both geographically and temporally, lead to changes in basic fares. Depending on market conditions, consumers may confront fares which are higher than basic fares (1.0x). For instance, 1.2x, which would yield a fare 20% higher than the basic fare. The authors use detailed information about how consumers react to surge pricing. In particular, they can observe whether consumers accept or reject rides when prices increase. This data on consumer behavior is key to their demand curve estimation strategy.

The paper assumes that although UberX prices are not random, as they depend on local demand and supply conditions, there is nevertheless a random component from the consumer perspective.[16] This allows the authors to use a regression discontinuity design analysis to estimate local demand elasticities using a whole range of surge prices for Chicago, Los Angeles, New York City and San Francisco. Using these elasticity estimates, they calculate that the consumer surplus from UberX in those cities amounted to 2.88 billion US$ in 2015. Assuming a proportional rule between consumer surplus and gross bookings, the article extrapolates the results to the United States and finds an estimate of 6.76 billion US$ in consumer surplus from UberX for the United States economy.

IV. MISALIGNMENT BETWEEN EXISTING REGULATIONS AND THE GIG ECONOMY

Peer-to-peer markets have an enormous potential to increase welfare and benefit society. Unfortunately, existing regulations often hamper the development of peer-to-peer markets across sectors in the economy. In Spain and elsewhere, gig economy business models are confronting several obstacles to enter markets and compete.

The taxi industry is an interesting case study. Services provided by Uber are facing severe restrictions, or even straight bans, across cities and countries throughout the world. Spain is a formidable example. UberX is banned in the

[15] The dataset has approximately 50 million consumer observations (sessions) over the first 24 weeks of 2015 in the cities of Chicago, Los Angeles, New York and San Francisco.

[16] UberX calculates its prices to an arbitrary number of decimal number. For example, 1.249x or 1.251x, which reflect similar demand and supply conditions. However, UberX makes consumers face discrete prices for the above prices. For example, 1.2x in the former case and 1.3x in the latter.

country by a court ruling[17] since the end of 2014. Currently, Uber provides services in Spain, although only in the cities of Madrid and –since March 2018– Barcelona.[18] However, it confronts several regulatory restrictions –at national, regional and local level– that significantly hinder its ability to compete.

As the Spanish case is a telling one, this section provides a brief economic overview of the main restrictions on competition in the Spanish taxi service regulations and evaluates these restrictions from the standpoint of efficiency and welfare. In addition, the next section (Section V) presents the consumer welfare loss arising from these restrictions.

Taxis are relatively small vehicles that provide point-to-point transportation services in urban areas (OECD, 2007). These services have traditionally been provided in three market segments: street hailing, taxi ranks and pre-booked services. Street hailing and taxi ranks are more predominant in high-density urban areas, while pre-booked services have traditionally been more present in outer and suburban urban areas and rural areas. Regulations across countries make a distinction between taxis and private hire vehicles, even though the service provided by both types of vehicles is identical from an economic standpoint. In Spain, existing regulations allow private hire vehicles to provide services in the pre-booked market segment only. Services provided by private hire vehicles are banned both in taxi ranks and in the hail market.

During the last decade, technological change has disrupted the taxi industry across countries, affecting not only how this service is provided but also the organization of traditional market segments. New applications such as Uber enable passengers using a smartphone to locate a vehicle in an urban area, arrange its service quickly and conveniently, have prior information about the vehicle and the driver, get the dispatch in a very short period of time, track the arrival of the vehicle in a map, estimate the likely charge for the ride and pay electronically through the application. E-hailing through smartphones has transformed the taxi industry forever, combining elements from both classic street hailing –the service is relatively immediate– and pre-booked services –as the taxi is not hailed in the street but arranged through the application. E-hailing has emerged as a substitute for traditional street hailing, intensifying competition (FTC, 2013).

[17] Uber entered the Spanish market in 2014 under the brand UberPop providing a similar service to the one provided by UberX in most United States cities. UberPop was banned in December 2014 by a Spanish court ruling.

[18] Currently, Uber provides services in these cities under the brand "UberX". Despite the name, the model substantially differs from the UberX model in the United States. Drivers must be private hire licensed drivers and the service is similar to a high-quality Uber service connecting passengers with authorized licensed drivers. In Spain, licenses for private hire vehicles are subject to strict quantity restrictions. Hence, these restrictions considerably hinder Uber's ability to compete.

As in many other countries, the Spanish taxi industry has been regulated as a legal monopoly for decades. It still is. Quantity and price restrictions constitute the main pillars of this monopoly regime. First, the Spanish public administrations –namely the local bodies– cap the total number of taxis that are allowed to provide services in an urban area, either a municipality or a broader metropolitan area in certain cases. In the last two decades, the number of licenses has barely changed in Spain. Entry is blocked. Second, the Spanish public administration –namely the local bodies– regulate taxi fares, which are generally fixed. Therefore, there is no price competition in the taxi industry.

Existing regulations foresee that taxi operators may coordinate through their associations to make proposals to the public administrations on fares, licenses and other elements set out in the regulation. Hence, the current framework includes all the ingredients of a well-functioning economic cartel: (i) quantity restrictions, (ii) ban on price competition and (iii) operators' coordination. Indeed, the Spanish taxi industry operates as a legal cartel. The quantitative estimates reviewed in the next section confirm that the effects of this cartel are the ones predicted by basic economic theory: higher prices, less output and a welfare loss.

Private hire vehicles' services are also heavily regulated. First, regulations prevent these vehicles from providing services both in taxi ranks and in the street hailing segment. Second, the supply of private hire vehicles is subject to quantity restrictions. The maximum number of private hire vehicles that are allowed to operate is pegged –through a ratio[19]– to the number of taxis. Therefore, entry is significantly blocked. Even though private hire vehicles have freedom to set their prices –their fares are not regulated– the quantity restriction is so stringent that their impact in terms of market discipline is insignificant. Private hire vehicles do not constitute a competitive constraint in the behavior of the taxi cartel. In addition, both taxis and private hire vehicles are subject to quality regulation, and they face additional restrictions, such as geographical restrictions to service provision. Private hire vehicles confront more stringent requirements than taxis in terms of fleet size, among others. Finally, private hire vehicles are not allowed to circulate on the bus/taxi lanes.

The Spanish regulatory framework has given rise to a dual regulatory regime between taxis and private hire vehicles. Despite the fact that taxis and private hire vehicles provide an identical economic service, they are subject to different regulatory provisions that shape a dual regulatory regime. This dual

[19] The current national regulatory framework in private hire vehicles services allows the regional public administrations in Spain –private hire vehicles' licenses are regional– to deny any new authorization of private hire vehicles in case their total number in the region exceeds the ratio of 1 private hire vehicle for 30 taxis. This quantitative restriction hinders the entry of private hire vehicles.

regime exists for one reason only: the aim of the Spanish public administrations to shield the Spanish taxi cartel from competition. For instance, as private hire vehicles have freedom to set their price, the regulator limits the supply of these vehicles through a quantity restriction. If this quantity restriction were removed, the monopoly regime in taxi services would be basically demolished. Other restrictions faced by private hire vehicles have arisen for the same reason. Hence, restrictions on competition included in private hire vehicle regulations constitute a corollary of inefficient regulation: they exist only to protect an inefficient and unjustified monopoly regime in taxi services.

Existing regulations in taxi and private hire vehicles services hinder the entry of gig economy business models. Quantity restrictions cap the number of licensed vehicles, either taxis or private hire vehicles. Hence, the entry of new licensed vehicles is basically blocked. Providing taxi or private hire transport services without an administrative licensed vehicle is illegal according to the Spanish public administrations in the transport sector. Therefore, the vehicles of a platform like UberX (United States model) are not allowed to operate in Spain. A service like UberPool –the Uber service that allows different passengers to share a ride– is also banned, as existing regulations ban car sharing in either taxi or private hire vehicles services. Other types of models, such as UberX (Spain model) or the one used by Cabify[20] are currently providing service in some cities. However, vehicle licenses for private hire vehicles are so restricted that Uber or Cabify platforms lack a critical mass of vehicles to compete. The regulatory misalignment is clear in other cases. A platform like Uber aims to compete on price and rely on dynamic pricing formulas, such as surge pricing. However, taxi regulations are at odds with this model: fares are regulated and price competition is banned. In addition, new business models resort to evaluation systems and reputational mechanisms to ensure quality. However, existing taxi and private hire vehicles regulations aim to solve quality problems with traditional ex-ante regulatory requirements. In sum, existing regulations block the entry of gig economy business models in taxi-type services and are misaligned with them.

The monopoly regime in taxi services has a negative impact on welfare. Quantity restrictions hinder rivalry and competition and have many adverse consequences. First, they lower vehicle availability. This is detrimental for all consumers, but especially for those with reduced mobility, such as persons with a disability or elder people. The empirical evidence confirms that quantity restrictions reduce vehicle availability, create scarcity and restrain consumer choice. In the United Kingdom, the areas with quantity restrictions had fewer taxis relative to population size than the areas without quantity restrictions (OFT, 2003). In Ireland, the number of taxis increased by 272% five years after the

[20] Cabify is a Spanish platform that was founded in 2011 and shares some features with Uber.

quantity restrictions were removed in 2000 (Commission for Taxi Regulation, 2009). In New Zealand, taxis increased by 160% from 1989 –when quantity restrictions were lifted– to 1994 (Bekken, 2006).

Second, quantity restrictions increase waiting times for customers. In addition, they provide lower incentives for taxi operators to improve quality and innovate. The empirical evidence suggests that waiting times decrease and quality improves when quantity restrictions are lifted. For instance, in the United Kingdom waiting times were 2%-7% lower in the areas without quantity restrictions than in the areas where quantity was restricted, all other things equal (OFT, 2003). Lower waiting times lead to valuable time savings for consumers. In addition, recent entry of platforms like Uber or Lyft in several cities has led to higher vehicle availability, lower waiting times, better service and higher innovation, valuable for consumers (Canada Competition Bureau, 2015).

Third, quantity restrictions distort consumers' choice between different transport options (OFT, 2003). This distortion is welfare detrimental because consumers end up choosing options that they value less than taxi services. In the United Kingdom, 15% of consumers from areas with quantity restrictions considered that higher waiting times were the main reason for using transport options other than taxis (OFT, 2003). Moreover, the market share of taxis in the field of transportation services shows a declining trend in many countries (OECD, 2007). A restrictive regulatory framework which has increased prices and waiting times has contributed to this long-term trend.

Fourth, quantity restrictions create an artificial scarcity of taxis, which is reflected in the value of a taxi license in the secondary market. Moreover, the artificial value of taxi licenses affects regulated fares in an indirect fashion, as fares are usually set in a cost-plus fashion. Hence, dynamically, this leads to a vicious circle that contributes to further increases in regulated fares and license values (OECD, 2007).

In Spain, in the year 2015 the value of a taxi license in the secondary market amounted to 220,271 euros in San Sebastian, 215,000 euros in Santander, 205,957 euros in Palma de Mallorca, 190,000 euros in Segovia, 142,254 euros in Madrid or 139,750 euros in Toledo.[21] Furthermore, scarcity increases

[21] Source: *Informe económico sobre las restricciones a la competencia incluidas en el Real Decreto 1057/2015 y en la Orden FOM/2799/2015, en materia de vehículos de alquiler con conductor,* Economic Analysis Unit of the CNMC, 8 June 2016 (CNMC, 2016c). It is worth emphasizing that these are values in the secondary market and that there is a considerable difference between this value and the administrative fee charged for issuing the license. In 2012, the administrative fee for issuing a license in the city of Cordoba was 457.13 euros while the value of a taxi license amounted to 102,102 euros (Source: *Informe económico sobre los límites cuantitativos y las restricciones a la competencia en precios en el sector del taxi de la ciudad de Córdoba,* Economic Analysis Unit of the CNMC, 15 January 2016 (CNMC, 2016a).

over time, because the procedures used by public administrations to control supply are subject to the influence and the lobbying actions of taxi associations. Public administrations, in Spain and elsewhere, respond to lobbying efforts by restricting supply,[22] which leads to increasing rents for incumbents. Regulatory capture (Stigler, 1971; Peltzman, 1976) is very common in the taxi industry (OECD, 2007; Fingleton, Evans and Hogan, 1997). Lobbying efforts and regulatory capture are the reasons why the number of taxi licenses does not respond to changes in demand and supply conditions. In Spain, the number of taxi licenses has barely changed in the last two decades, and in many of its regions and municipalities, it has decreased over the last decades. For instance, in the Barcelona Metropolitan Area no new taxi licenses have been issued since 1980, even though population and GDP have increased substantially. The city of Bilbao has not issued new taxi licenses since 1978,[23] four decades ago.

Due to lobbying efforts, regulatory capture and increasing scarcity, the value of taxi licenses exhibits an increasing trend through time. Furthermore, over time, taxi licenses become an asset yielding a high return, even higher than the return on stocks. For instance, between 1987 and 2016, the annual compound growth rate of the value of a taxi license in the Barcelona Metropolitan Area was 6.4%, compared to a rate of 4.2% in the case of the IBEX-35 index, the main Spanish stocks reference index. In absolute terms, the value of a taxi license has increased by 503.7% between 1987 and 2016, compared to 233.7% in the case of the IBEX-35 index. Between 2001 and 2016, returns have been considerably higher in the case of a taxi license (see Figure 1). In a competitive market, a license would have no value. Hence, this dynamic shows the extraordinary degree of inefficiency arising from the monopoly regime in taxi services.

Quantity restrictions in taxi services lack a justification on welfare grounds. OECD (2007) reviews many of the possible justifications for restricting supply. Among the possible arguments in favor of quantity restrictions, the most promising one *a priori* would be the one related to congestion and pollution. Advocates of this view argue that free entry would lead to an excessive number of taxis, leading to congestion and pollution externalities (Shreiber, 1975). However, one shortcoming of this argument is that congestion and pollution

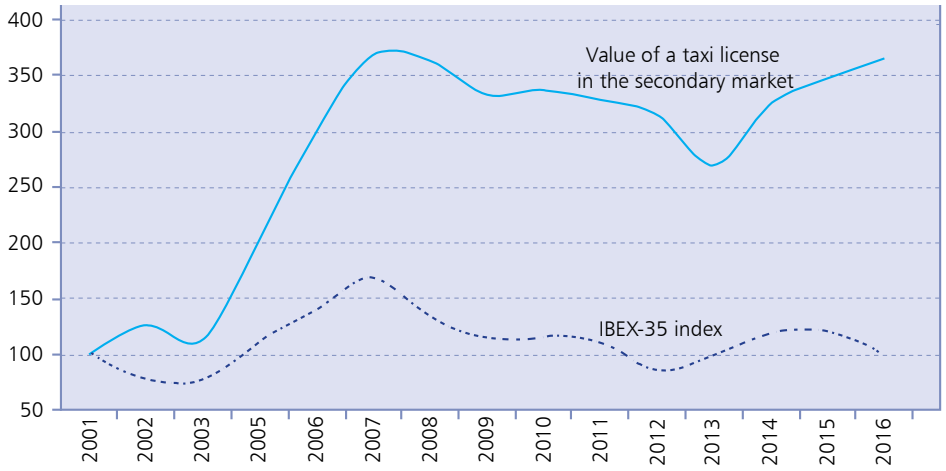
[22] In particular, the Spanish public administrations may respond to lobbying efforts by not issuing new licenses or by removing licenses from the market using taxpayers' money. For instance, in May 2017 the city of Santa Cruz de Tenerife and the regional authority Cabildo de Tenerife agreed to allocate an aggregate amount of 4 million euros to remove taxi licenses from the market. http://www.tenerife.es/portalcabtfe/es/el-cabildo/portal-de-transparencia/como-lo-hacemos/convenios-encomiendas/listado-de-convenios-y-encomiendas/126-area-de-presidencia/6304-convenio-de-colaboracion-entre-el-excmo-cabildo-insular-de-tenerife-y-el-excmo-ayuntamiento-de-santa-cruz-para-el-rescate-de-licencias-de-auto-taxi

[23] Source: Spanish Statistical Institute (INE) and *Historia del Taxi de Bilbao,* research document sponsored by Radio Taxi Bilbao and written by Julio Pérez, which can be downloaded here: http://docplayer.es/4577856-Historia-del-taxi-de-bilbao.html

FIGURE 1

VALUE OF A TAXI LICENSE OF THE BARCELONA METROPOLITAN AREA IN THE SECONDARY MARKET AND THE IBEX-35 INDEX (SPANISH STOCKS REFERENCE INDEX), 2001-2016*

(2001 = base 100)



Note: *A figure similar to this one and built using identical data was also inserted in *Informe Económico sobre el Decreto 314/2016, relativo a la actividad de mediación en los servicios de taxi en Cataluña,* Economic Analysis Unit of the CNMC, 15 June 2017 (CNMC, 2017).

Sources: Elaborated from (i) data on the value of taxi licenses in the Barcelona Metropolitan Area from the Taxi Metropolitan Institute of the Barcelona Metropolitan Area and (ii) data on the Ibex-35 Index from Banco de España. Data on the value of a taxi license are not available for the year 2004.

could be lower with free entry. Free entry would increase taxi availability and reduce waiting times. Therefore, it could also trigger a substitution effect between taxis and private vehicles, leading to a lower degree of private vehicle utilization and a lower level of congestion and pollution originating from private vehicles (OECD, 2007; Frankena and Pautler, 1984). This effect would be more intense in a scenario where price restrictions were also removed. Moreover, in the current technological context, the entry of gig economy business would additionally contribute to lowering the degree of congestion and pollution in urban areas. Li, Hong and Zhang (2017) provide evidence for the United States suggesting that ride-sharing services such as Uber decrease traffic congestion time, congestion costs, and excessive fuel consumption. The literature also finds that ride-sharing platforms contribute to reducing car ownership and trigger a substitution effect of ride-sharing platforms for vehicle use (Hampshire *et al.,* 2017).

In any case, in the unlikely event that lifting quantity restrictions led to an increase in urban congestion and pollution, the most adequate way to tackle this issue is through a general policy measure aiming to correct congestion and pollution externalities arising from vehicle use. After all, congestion and pollution arise from any type of vehicle use, not only taxis or private hire vehicles. In addition, this type of vehicles account for a very small part of the total stock of vehicles in an urban area (Frankena and Pautler, 1984). Hence, resorting to sectoral regulation in taxi services is not an efficient way to address congestion and pollution externalities. Moreover, quantity restrictions constitute a blunt measure to correct congestion and pollution externalities related to vehicle use. Price mechanisms such as tolls constitute a superior measure in terms of efficiency (Brueckner, 2011).

Opponents of free entry have additional arguments (OECD, 2007), even less powerful than the one relating to congestion and pollution externalities. On the one hand, they argue that free entry would reduce vehicle capacity utilization, lowering average vehicle productivity and leading to price increases. Nevertheless, at best, this argument could be taken into consideration in the cases where fixed costs are relevant and sunk. However, this is not the case in the taxi industry. In addition, the argument is inconsistent with empirical evidence. In the cases where quantity restrictions have been lifted, such as New Zealand, prices have decreased, contrary to what the proponents of quantity restrictions argue that would have happened. On the other hand, advocates of quantity restrictions argue that removing quantity restrictions would increase competition and incentivize undertakings to reduce quality. However, competition also fosters quality, as taxi operators compete in quality, especially in the current technological environment.[24] Moreover, quality is separable from quantity. If there is any type of market failure linked to quality that the market cannot solve, regulation can in principle tackle quality related issues. Restricting quantity is neither necessary nor adequate to correct market failures linked to informational problems and quality provision.

Price restrictions constitute the second main pillar of the monopoly regime in taxi services, both in Spain and in other countries. In Spain, taxi fares are generally regulated and fixed. They constrain price competition and hinder the emergence of efficient prices in taxi services. In particular, they hamper the development of dynamic pricing formulas used by modern peer-to-peer platforms that respond to changes in demand and supply conditions. Dynamic pricing is an efficient way to match supply and demand in urban transportation markets.

[24] Platforms such as Lyft or Uber solve or mitigate informational problems through external regulations –such as background checks on drivers– and evaluation systems and reputational mechanisms, enabling intense competition in quality. In the past, competition in quality was relatively more constrained as information problems in the taxi market were more acute.

Traditionally, two factors have been relevant when assessing the possibility of price competition in the taxi industry: search costs and information costs (Shreiber, 1975). Search costs may hamper consumers' ability to find suppliers and compare prices. If relevant, they may confer market power (Diamond, 1971). In taxi services, search costs have traditionally been linked to the costs of finding a taxi, waiting for another taxi when one had already been hailed or the risk of losing a relatively good quote (Fingleton, Evans and Hogan, 1997). Information costs may affect certain types of consumers, such as infrequent consumers or tourists, among others. Information and search costs could provide locational market power to a taxi hailed in the street.

Search and information costs have traditionally permeated the arguments favoring regulated fares in the taxi industry, even though part of the economic literature has argued that if quantity restrictions were lifted, search costs would not be extremely relevant and brands/fleets could tackle informational problems (Coffman, 1977). The empirical evidence suggests that the latter idea deserves merit. Price restrictions and quantity restrictions were removed in New Zealand in 1989. Prices dropped between 15% and 25% in real terms in urban areas (Bekken, 2006). With a well-designed framework, price competition has always been possible and desirable in the taxi industry. New Zealand opted to include an obligation to make tariffs publicly available –inside and outside the vehicle– but without restricting freedom to set prices (PC, 1999). New Zealand chose an efficient way to allow for price competition while at the same time correcting for the possible locational market power of a taxi driver in certain circumstances –for instance, during a storm.

Despite the traditional concerns regarding the possibility of price competition in the taxi industry, recent technological progress has substantially increased the scope for price competition and further weakened the arguments of those in favor of price controls. Technological progress has lowered search and information costs an allowed for intense and vigorous price competition. Through the innovative services of ride-sharing platforms, users can request a ride having prior information on prices, the expected cost for the ride, vehicle and driver information and expected waiting and travel times. Payments can be done through the platform –there is no need to use a cash or a credit card– and after the ride the user receives an email with the receipt and information about the journey. Improvements for users are evident. In addition, as mentioned before, peer-to-peer markets resort to valuation systems and reputational mechanisms to incentivize drivers to provide services ensuring quality. For all these reasons, transaction and search costs are minor and users have adequate information on prices and quality when arranging the service, which is immediate. In this context, price competition can be vigorous and intense. Moreover, thanks to technological progress, dynamic pricing schemes are now possible in the taxi

industry. These allow for more efficient pricing and contribute to increasing both supply elasticity and the availability of service.

In the current technological environment, regulated fares –either fixed or maximum– are not justified on welfare grounds. Regulated maximum fares could be justified in a scenario with quantity restrictions in place yielding market power to incumbent operators. However, as quantity restrictions are not justified in terms of welfare, there is no justification for regulating fares.

Taxi and private hire regulations in Spain incorporate other restrictions on competition, such as restrictions related to safety and quality assurance –based on standards on both vehicles and drivers–, geographical restrictions to service provision,[25] or bans on car pooling, among others. None of these restrictions is justified on welfare grounds in its current configuration. First, certain requirements on vehicles and drivers related to quality assurance could be justified in the past, given informational problems affecting consumers' ability to ascertain service characteristics. However, modern peer-to-peer markets solve informational problems very effectively. Therefore, market discipline and verification and reputational mechanisms, together with general regulations on vehicle safety, are in the majority of cases sufficient to ensure quality and safety issues. Regulated taximeters or regulatory requirements on specific inspections of vehicles,[26] among others, are not justified. Neither regulatory requirements on drivers such as public examinations, dress codes or an administrative authorization to become a driver. Second, geographical restrictions to service provision constrain competition and lead to significant inefficiencies. In particular, they hamper the generation of scale and scope economies in service provision and prevent further reductions in price. They are not necessary to correct any type of market failure. Hence, they lack a justification and should be removed. Third, car pooling is banned. If there are any informational problems in this case, current available technology is capable of solving them. Car pooling for taxi-type services already exists in many cities such as London, Mexico City or New York City. Moreover, pooling exists in Spain for different modes of transportation –such as the subway– and in a peer-to-peer market such as Blablacar, which provides services in interurban transportation. Hence, car pooling should also be allowed in taxi type services, either taxis or private hire vehicles.

In sum, the current taxi monopoly regime lacks a justification from the standpoint of welfare and the public interest. In particular, quantity restrictions,

[25] In essence, these restrictions ban or substantially hinder the ability of taxi operators or private hire vehicles to fully provide services outside the jurisdiction where they obtained their license.

[26] Those that go beyond the requirements that any private vehicle must fulfill according to the legal framework.

price restrictions and geographical limitations to service provision should all be removed, so as to establish a system of free entry and price competition across the Spanish territory. Regulations should not make a distinction between taxis and private hire vehicles as they provide the same economic service.

The current regulatory framework in taxi services is just one particular example of an unjustified regulation which is hindering competition and blocking the entry of gig economy business models. In Spain and elsewhere there are many other examples, such as the lodging industry, the interurban transport industry or the finance industry, among others. The next section presents some estimates of how harmful unjustified regulations can be. To continue with the narrative, the monopoly regime in the Spanish taxi industry is used as an illustration.

V. WELFARE COSTS OF UNJUSTIFIED REGULATIONS

Unjustified anticompetitive regulations harm consumers and society as a whole. They do so because they hinder competition and innovation, lower productivity and ultimately lead to lower economic growth and living standards. If they are not removed from the regulatory framework, consumers and society will continue to bear the costs. This section presents some quantitative estimates of the consumer welfare loss arising from the monopoly regime in the Spanish taxi industry.

The economic literature has quantified the welfare costs arising from the monopoly regime in taxi services for many cities. Swan (1979) explores the effect of entry restrictions in the taxi market of Canberra (Australia). He concludes that removing entry restrictions would decrease prices for taxi services by 13.6%. Taylor (1989) quantifies the welfare cost of the restrictions on competition in the taxi market of Toronto (Canada). He finds that taxi regulation reduced the quantity of taxis by 730 and increased prices by 33% compared to a counterfactual competitive scenario. The consumer welfare loss in Toronto amounted to 39.2 million dollars in 1987. Gaunt and Black (1996) undertake a similar quantitative exercise for the city of Brisbane (Australia). They find that entry and price restrictions lowered the number of taxis by 228 and increased prices by 15.6%. The consumer welfare loss amounted to 20.6 million Australian dollars per year.

In the case of Spain, the Economic Analysis Unit of the Spanish Competition and Markets Commission (SAE) has quantified the consumer welfare loss arising from the monopoly regime in taxi services in several cities, metropolitan

areas and the Spanish economy as a whole.[27] CNMC (2015), CNMC (2016a), CNMC (2016c) and CNMC (2017) refer to quantifications undertaken in several economic reports elaborated for jurisdictional challenge appeals of the CNMC against taxi or private hire regulations at local, regional or national level.[28] Silos (2017) compiles many of those quantifications and extends the quantification methodology to estimate the consumer welfare loss caused by the taxi monopoly in other Spanish cities and urban areas. All the quantifications are on a per year basis. The base year varies across quantifications because of (i) data availability in each case and (ii) the time when the quantifications were undertaken.

In the case of Malaga (CNMC, 2015), the annual consumer welfare loss arising from entry and price restrictions in taxi services amounted to 4.4 million euros in 2012. The price overcharge compared to a counterfactual scenario where entry and price restrictions were removed was 10%-11%. In the case of Cordoba (CNMC, 2016a), the annual consumer welfare loss was 2.5 million euros in 2012. In the case of the Barcelona Metropolitan Area (CNMC, 2017), the annual consumer welfare loss amounted to 61.4 million euros in 2016 and a price overcharge of 12.3%. At national level, CNMC (2016c) estimates that the monopoly regime in taxi services increased prices by 13.1% and generated an annual consumer welfare loss of 324.3 million euros in 2013. Silos (2017) provides additional quantifications for other cities in Spain in 2015.[29] For instance, the annual consumer welfare loss amounted to 96.2 million euros in the Area for Common Taxi Services Provision of Madrid, 11 million euros in Palma de Mallorca, 2.9 million euros in San Sebastián, 2.5 million euros in Granada or 2.4 million euros in Santander. Price overcharges estimates included in Silos (2017) fall within the interval of 11% –27% across cities and areas.

All the quantifications are conservative estimates for two main reasons. First, they do not take into account the productive and dynamic inefficiencies that arise in a non-competitive environment. Second, they do not take into account the welfare loss arising from higher waiting times for consumers in a non-competitive environment, where the number of taxis is lower. All the quantifications follow an approach based on Gaunt and Black (1996). As an example, a more detailed explanation of the quantification undertaken for the Spanish economy as a whole is provided below.[30]

---

[27] These quantifications constitute a contribution to the empirical literature as they were the first quantifications undertaken to estimate the welfare loss arising from the monopoly regime in taxi services in Spain. They were all undertaken while the author of this article was Head of Economic Analysis of the Spanish Competition and Markets Commission (CNMC).

[28] Most of these appeals are still pending. The end of this section briefly discusses the only one which has already been decided at court. In this case, the court's ruling has been unfavorable to the CNMC's core claims.

[29] In two cases the reference year is 2016.

[30] This explanation basically follows the one provided in CNMC (2016c) and Silos (2017).

The quantification starts out from the value of a taxi license (medallion) in the secondary market. The value of a license reflects the discounted value of the future stream of supra-competitive rents that arise in the monopoly environment. The value of a taxi license transacted in the secondary market is calculated using data from the Spanish Tax Agency on secondary market transactions of taxi licenses. Using data from 41 Spanish provinces in 2013 and a total of 1,503 transactions, the estimated value of a license at national level amounted to 93,426.4 euros in the year 2013. Using the average interest rate over the period 2000-2013 of the Spanish Treasury's ten-year bond (4.6%), the annual monopolistic rent of a taxi license in the Spanish economy would amount to 4,297.6 euros. Given the number of taxi licenses in Spain in 2013 (70.808), the total value of supra-competitive rents in 2013 amounts to 304.3 million euros. This figure represents most of the consumer welfare loss arising from the monopoly in taxi services in Spain in 2013.

Starting from the total value of supra-competitive rents, the quantitative exercise estimates the price overcharge arising from the monopoly situation. According to Eurostat, the total revenues of taxi services in Spain amounted to 2,320.8 million euros in 2013. The weight of the total value of supra-competitive rents on that figure amounts to 13.1%. This figure represents a conservative estimate of how much prices would be reduced if entry and price restrictions were removed. Assuming a price demand elasticity of 1 in taxi services (see, for instance, Gaunt and Black, 1996; Taylor, 1989; Frankena and Pautler, 1984) and taking into account the price reduction of 13.1%, the number of taxi licenses in the Spanish economy would increase by 9.284 if entry and price restrictions were removed. This is a conservative threshold of the artificial scarcity in the number of taxi licenses created by the monopoly in taxi services.

With all the above data, it is possible to estimate the total consumer welfare loss arising from the monopoly regime in taxi services in Spain. First, the welfare loss for the quantity that consumers consume under a monopoly environment amounts to 304.3 million euros. Second, the consumer welfare loss due to the lower quantity consumed in the monopoly situation compared to the quantity consumed if entry and price restrictions were removed, amounts to 20.0 million euros. Adding up these two figures yields a total consumer welfare loss of 324.3 million euros in 2013, which is a conservative estimate of the annual consumer welfare loss under the current monopoly regime.

Hence, actions by the Spanish Government and regional and local public bodies intending to shield and protect the current monopoly regime in taxi services in Spain impose a welfare cost on consumers amounting to a conservative estimate of 324.3 million euros.

Despite all this evidence, the Spanish public administrations persist on maintaining the current monopoly regime in taxi services. For instance, the Spanish Government had the opportunity to remove the quantity restrictions and other restrictions in the private hire vehicle sector at the end of 2015. However, it decided to keep all the restrictions in the regulation (a Royal Decree), in particular the quantity restriction. Moreover, more recently, the Spanish government has decided to include the quantity restrictions and other restrictions in a Royal Legal Decree, so as to shield them even more, in a desperate and urgent move to protect the taxi monopoly.[31] Regional and local administrations are behaving in the same fashion.

The case of taxi services is just an example of what is also happening in other sectors.[32] In recent years, national, regional and local administrations in Spain are increasing unjustified regulatory restrictions on competition in the economy, seeking to hamper the capacity of gig economy business models to enter markets and compete.

Across sectors, regulators are behaving as they usually do when innovation thrives and disrupts markets. First, they are neglecting the benefits of entry and competition. Second, they are worried about the adverse effects of entry on incumbent players. Third, they are working to preserve the regulatory *statu quo,* opposing beneficial reform for society. Sadly, the sharing economy is yet another example of how vested interests, lobbying efforts and regulatory capture can block the path to pro-competitive reform, depriving society of higher levels of prosperity brought by modern technological progress.

But society deserves better. Technological progress and competitive markets are key to foster innovation, increase productivity and improve welfare

[31] After the Spanish Government decided to uphold the quantity restrictions and other restrictions in the Royal Decree of 2015, the CNMC challenged at court this Royal Decree. In fact, the aforementioned economic report quantifying the consumer welfare loss arising from the taxi monopoly in the Spanish economy was elaborated for the CNMC's jurisdictional challenge. As the High Court's ruling could lead to the removal of quantity restrictions and other restrictions in private hire vehicles, on 20 April 2018, when the ruling was imminent, the Spanish Government passed a Royal Legal Decree that included those restrictions. This Royal Legal Decree was validated by a large majority of the Spanish Parliament on 10 May 2018. The Spanish Government had two objectives. First, to influence the Spanish High Court, intending to make its ruling favorable to the Spanish Government's restrictive policy in urban transportation. Second, to protect even more the taxi monopoly from (i) a High Court's ruling favoring the removal of the aforementioned restrictions and (ii) another future jurisdictional challenge from the CNMC. As it is a Royal Legal Decree, the CNMC lacks powers to challenge it at court. The CNMC can only challenge regulations which rank below laws. Llobet (2018) highlights that this recent regulatory move by the Spanish Government constitutes a prototypical case of regulatory capture. In the end, on 4 June 2018, the Spanish High Court decided to uphold the quantity restrictions in private hire vehicles.

[32] See for instance CNMC (2016b) for the lodging industry or the interurban bus services, Maudes (2018) for the lodging industry or Llobet (2014a) and Llobet (2014b) for the banking industry and the lodging industry.

for all. Competition is a source of welfare. This is why competition should not be restricted unless there is market failure or public interest justifying it. Absent this justification, competition should prevail. Otherwise, society will lose. Policy makers should bear in mind that the gig economy solves or mitigates the very same market failures that allegedly motivated existing regulations in the past. And that as a result, many regulatory provisions intending to correct them become unnecessary and should be removed or modified. The gig economy constitutes an opportunity to improve how we regulate markets and improve welfare for all.

But policy makers already know that. In fact, the *true* challenge that the gig economy poses for policy makers is not to surrender to regulatory capture and avoid serving the interest of incumbent players by stifling healthy competition and innovation. So far, they have failed to succeed in that challenge. Let us hope that in the future, both in Spain and elsewhere, regulators finally decide to reform existing regulations so as to welcome the development and flourishment of the gig economy by promoting a pro-competitive and efficient regulatory framework.

VI. CONCLUSIONS

In the last decade, peer-to-peer markets have flourished in many sectors of the economy, ranging from rental accommodation to transport services, professional services, financial services or household tasks, among others. These markets help buyers and sellers find each other and allow them to share assets or resources through the Internet. They are often referred to as the "gig", "sharing" or "collaborative" economy and they constitute one of the latest byproducts of modern technological progress.

Sharing underutilized assets by individuals was certainly possible before the emergence of modern peer-to-peer markets. However, exchanges were relatively scarce. High transaction costs and informational problems usually led to risky and costly transactions, making them very infrequent and often restricted to exchanges with relatives or friends. Taking advantage of technology, modern peer-to-peer markets overcome these problems, lowering transaction costs and ensuring trustworthy transactions. Trust is a crucial factor to create the network effect between buyers and sellers and explains the success and popularity that peer to peer markets have achieved.

Peer-to-peer markets lower entry barriers, enhance competition and disrupt markets. Its emergence has already transformed many industries forever. The case of taxi services is a paramount example. For decades, this industry has

been regulated as a legal monopoly across countries. In the last ten years, platforms such as Lyft or Uber have succeeded in using new technologies to match drives and passengers in effective and innovative ways, increasing efficiency, expanding supply and fostering competition in terms of price, quality and variety. Where these platforms are providing service with an effective ability to compete, they have intensified competition and massively attracted drivers and passengers. In certain cases, they have fully eroded the monopoly position of incumbent operators. Markets for rental accommodation, financial services, professional services, household tasks, touristic guides or even babysitting are also being disrupted by the gig economy.

The potential benefits of the gig economy for society are substantial. Improvements in efficiency and increased competition lead to lower prices, higher quality, more variety, increased innovation, higher productivity and, ultimately, more economic growth and higher living standards. In sum, better lives for all. Unfortunately, existing regulations are hampering the development of peer-to-peer markets across sectors. Again, the taxi industry is an interesting case. Services provided by a company like Uber are facing severe restrictions, or even straight bans, across cities and countries throughout the world. Spain constitutes a formidable example. Incumbent players in the taxi industry are reacting strongly against innovative entrants and lobbying to maintain or reinforce existing regulations so as to stifle healthy competition and protect the monopoly regime in taxi services.

When existing regulations were first enacted, regulators often advocated them on welfare grounds. Their intended purpose, so they said, was to correct market failures and enhance welfare in comparison to the outcome which would have prevailed in an unregulated market. Although many sectoral regulatory provisions were just a byproduct of anticompetitive rent seeking –quantity restrictions in the taxi industry are a classic example– the fact is that peer-to-peer markets solve or mitigate through innovative ways the very same market failures that allegedly motivated existing regulations in the first place. Their success and popularity are precisely rooted in their capacity to make markets work in an efficient fashion, by lowering search costs and solving informational problems. As peer-to-peer markets contribute to solving or mitigating market failures, many regulatory provisions intending to correct them become unjustified and should be removed or modified. The gig economy gives society the opportunity to improve how it regulates markets and improve welfare for all.

Unjustified traditional regulations harm consumers and society as a whole. They do so because they hamper competition, lower productivity and ultimately lead to lower economic growth and living standards. Again, a telling case is the welfare loss arising from the monopoly regime in taxi services, which lacks a justification

on welfare grounds. In the case of the Spanish economy as a whole, consumers pay an overcharge of 13.1% for these services. The annual consumer welfare loss caused by this monopoly regime amounts to 324.3 million euros per year. And these are conservative estimates.

Despite the potential benefits of the gig economy, reforming existing regulations is proving very hard. Regulators are behaving as they usually do when innovation thrives and disrupts markets. They are worried about the adverse effects of entry on incumbent players and working to preserve the regulatory *statu quo.* Regulatory capture is blocking the path to beneficial reform for society.

But society deserves better. Let us hope that in the future, regulators welcome the development and flourishment of the gig economy by promoting a pro-competitive and efficient regulatory framework which would increase welfare for all.

BIBLIOGRAPHY

Bagchi, S. (2017), "A Tale of Two Cities: An Examination of Medallion Prices in New York and Chicago," mimeo.

Bekken, J. T. (2006), "Experiences with Regulatory Changes of the Taxi Industry," *9th Conference on Competition and Ownership in Land Transport.*

Belleflame, P. (2017), "Les plateformes de l'économie collaborative: fonctionnement et enjeux," in *La consommation collaborative. Enjeux et défis de la nouvelle société du partage,* Decrop, A. (editor), De Boeck.

Bialik, C.; Flowers, A.; Fischer-Baum, R., and D. Mehta (2015), "Uber Is Serving New York's Outer Boroughs More Than Taxis Are," *FiveThirtyEight,* August.

Brodeur, A., and K. Nield (2016), Has Uber Made It Easier to Get a Ride in the Rain?, *IZA Discussion Paper Series,* 9986.

Brueckner, J. K. (2011), *Lectures on Urban Economics,* MIT Press.

Caillaud, B., and B. Jullien (2003): "Chicken & Egg: Competition Among Intermediation Service Providers," *RAND Journal of Economics,* Vol. 34 (2): 309-328.

Canada Competition Bureau (2015), Modernizing Regulation in the Canadian Taxi Industry, *White Paper.*

City of Ottawa (2015), *Ottawa Taxi and Limousine Regulations and Services Review – Customer Experience,* research undertaken by Core Strategies for the City of Ottawa.

CNMC – Comisión Nacional de los Mercados y la Competencia (2015), *Informe económico sobre los límites cuantitativos y las restricciones a la competencia en precios en el sector del taxi de la ciudad de Málaga,* Economic Analysis Unit of the CNMC, December 22.

— (2016a), *Informe económico sobre los límites cuantitativos y las restricciones a la competencia en precios en el sector del taxi de la ciudad de Córdoba,* Economic Analysis Unit of the, January 15.

— (2016b), *Resultados preliminares. E/CNMC/004/15 Estudio sobre los nuevos modelos de prestación de servicios y la economía colaborativa.*

— (2016c), *Informe económico sobre las restricciones a la competencia incluidas en el Real Decreto 1057/2015 y en la Orden FOM/2799/2015, en materia de vehículos de alquiler con conductor,* Economic Analysis Unit of the, June 8.

— (2017), *Informe Económico sobre el Decreto 314/2016, relativo a la actividad de mediación en los servicios de taxi en Cataluña,* Economic Analysis Unit of the, June 15.

Coffman, R. B. (1977), "The Economic Reasons for Price and Entry Regulation of Taxicabs: A Comment," *Journal of Transport Economics and Policy,* Vol. 11 (3): 288-297.

Cohen, P.; Hahn, R.; Hall, J.; Levitt, S., and R. Metcalfe (2016), Using Big Data to Estimate Consumer Surplus: The Case of Uber, *NBER Working Paper,* 22627 September.

Commission for Taxi Regulation (2009), *Economic Review of the Small Public Service Vehicle Industry,* Irish Commission for Taxi Regulation, elaborated by Goodboy Economic Consultants.

Cramer, J., and A. B. Krueger (2016), "Disruptive Change in the Taxi Business: The Case of Uber," *American Economic Review: Papers & Proceedings,* Vol. 106 (5): 177–182.

Diamond, P. A. (1971), "A Model of Price Adjustment," *Journal of Economic Theory,* Vol. 3 (2): 156-168.

Einav, L. (2014), "The Economics of Peer-to-Peer Internet Markets," presentation given at a conference held at the Federal Trade Commission, Washington D.C., October, 2014.

— (2015), "The Economics of Peer-to-Peer Internet Markets," presentation given at *The "Sharing" Economy,* conference held at the Federal Trade Commission, Washington D.C., June.

Einav. E.; Farronato, C., and J. Levin (2016), "Peer-to-Peer Markets," *Annual Review of Economics,* 8: 615-635.

Fingleton, J.; Evans, J., and O. Hogan (1997), *The Dublin Taxi Market: Re-regulate or Stay Queuing?,* Department of Economics, Trinity College, Dublin.

Fraiberger, S. P., and A. Sundararajan (2017), "Peer-to-Peer Rental Markets in the Sharing Economy," *NYU Stern School of Business Research Paper.*

Frankena, M. W., and P. A. Pautler (1984), *An Economic Analysis of Taxicab Regulation,* Staff Report of the Bureau of Economics of the Federal Trade Commission.

FTC – Federal Trade Commission (2013), *FTC Staff Comments Before the District of Columbia Taxicab Commission Concerning Proposed Rulemakings on Passenger Motor Vehicle Transportation Services.*

Gaunt C., and T. Black (1996), "The Economic Cost of Taxicab Regulation: The Case of Brisbane," *Economic Analysis and Policy,* 26(1): 45-58.

Hall, J. V., and A. B. Krueger (2017), "An Analysis of the Labor Market for Uber's Driver-Partners in the United States," *ILR Review.*

Hampshire, R.; Simek, C.; Fabusuyi, T.; Di, X., and X. Chen (2017), "Measuring the impact of an unanticipated suspension of ride-sourcing in Austin, Texas," mimeo.

Hicks, J. R. (1935), "Annual Survey of Economic Theory: The Theory of Monopoly," *Econometrica,* Vol. 3 (1): 1-20.

Horton, J. J., and R. J. Zeckhauser (2016), Owning, Using, Renting: Some Simple Economics of the 'Sharing Economy, *NBER Working Paper Series.*

Li, Z.; Hong, Y., and Z. Zhang (2017), "Do On-demand Ride-sharing Services Affect Traffic Congestion? Evidence from Uber Entry," mimeo.

Llobet, G. (2014a), "¿Y ahora toca regular el Crowdfunding?," *Nada es Gratis,* April.

— (2014b), "El legislador justiciero ataca de nuevo," *Nada es Gratis,* July.

— (2018), "La política del qué hay de lo mío," *Nada es Gratis,* April.

Maudes, A. (2018), "Short-Term Rentals in the Canary Islands. Regulations affecting the digital economy," *Medium,* January.

Maudes, A.; Sobrino, M., and P. Hinojo (2017), "Fundamentos Económicos de la Economía Colaborativa," in *Anuario de la Competencia 2016,* Lluís Cases (ed.), Marcial Pons: 167-186

OECD – Organization for Economic Co-operation and Development (2007), *Taxi Services: Competition and Regulation, OECD Competition Policy Roundtables.*

— (2017), *OECD Economic Survey of Australia.*

OFT – Office of Fair Trading (2003), *The regulation of licensed taxi and PHV services in the UK.*

PC – Productivity Commission (1999), *Regulation of the Taxi Industry,* Ausinfo, Canberra.

Peltzman, S. (1976), "Toward a More General Theory of Regulation," *Journal of Law and Economics,* vol. 19 (2): 211-240.

PWC - Price Waterhouse Coopers (2016), *Assessing the size and presence of the collaborative economy in Europe,* authors: Vaughan R., and R. Daverio. This paper was produced for European Commission's DG GROW.

Rochet, J. C., and J. Tirole (2003), "Platform Competition in Two-Sided Markets", *Journal of the European Economic Association,* Vol. 1 (4): 990-1,029.

— (2006), "Two-sided markets: a progress report," *RAND Journal of Economics,* Vol. 37 (3): 645-667.

SherpaShare (2016), "Uber trips are becoming longer and faster, but are they more profitable?," SherpaShare Blog, http://www.sherpashareblog.com.

Shreiber, C. (1975): "The Economic Reasons for Price and Entry Regulation of Taxicabs," *Journal of Transport Economics and Policy,* Vol. 9 (3): 268-279.

Silos, M. (2017), "Estimación del daño ocasionado por el régimen de monopolio en los servicios de taxi en España," *Documento de Trabajo en Política de Competencia y Regulación* N.º 001/2017, Comisión Nacional de los Mercados y la Competencia (CNMC), January 2017. https://www.cnmc.es/sites/default/files/editor_contenidos/Promocion/CNMC_001_2017.pdf

Silverstein, S. (2014), "These Animated Charts Tell You Everything About Uber Prices In 21 Cities," *Business Insider,* October.

Spence, M. (2015), "The Inexorable Logic of the Sharing Economy," *Project Syndicate,* September.

Stallibrass, D., and J. Fingleton (2016), *Disruptive innovation in Latin America and the Caribbean: Competition enforcement challenges and advocacy opportunities,* Latin American and Caribbean Competition Forum.

Stigler, G. J. (1971), "The theory of economic regulation," *Bell Journal of Economics,* Vol. 2 (1)1971: 3–21.

Sundararajan, A. (2014a), "Peer-to-Peer Businesses and the Sharing (Collaborative) Economy: Overview, Economic Effects and Regulatory Issues," Written testimony for the hearing titled, The Power of Connection: Peer-to-Peer Businesses, held by the Committee on Small Business of the United States House of Representatives, January 15th, 2014.

— (2014b): "Trusting the 'Sharing Economy' to Regulate Itself," *The New York Times,* March.

Swan, P L. (1979), "On Buying a Job: The Regulation of Taxicabs in Canberra," Centre for Independent Studies.

Taylor, D. W. (1989), "The Economic Effects of the Direct Regulation of the Taxicab Industry in Metropolitan Toronto," *Logistics and Transportation Review,* Vol. 25 (2): 169-182.

The Economist (2013), "The Rise of the Sharing Economy," March.

US (United States) Department of Transportation (2011), *2009 National Household Travel Survey,* June.

Vulkan, N.; Roth, A., and Z. Neeman (Editors) (2013), *The Handbook of Market Design,* Oxford University Press.

Zervas, G; Proserpio, D., and J. W. Byers (2017), "The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry," *Journal of Marketing Research,* Vol. 54 (5): 687–705.

ECONOMICS OF NEWS AGGREGATORS[1]

Doh-Shin JEON

Abstract

The success of news aggregators has generated a heated debate about whether news aggregators steal traffic from newspapers or increase traffic to newspapers. This survey article provides an overview of recent articles on news aggregators. After providing a simple theoretical framework, I first review empirical articles and then theoretical ones. While the empirical articles try to assess the effects on traffic to newspapers, the theoretical ones go beyond and try to analyze the effects on newspapers' incentive to invest in quality journalism. I conclude by raising some questions for future research.

Key words: News aggregators, newspapers, traffic.

JEL classification: L80, L82.

[1] I would like to thank Luis Abreu, Gerard Llobet and Nikrooz Nasr for the very useful comments.

I. INTRODUCTION

The traditional ad-based business model of newspapers has been in crisis because of declining revenues from newspaper advertising. According to Pew Research Center (2017), newspapers' revenues from advertising have fallen approximately 62% since 2000 in the U.S.: it was $48.67 billion in 2000 but $18.27 billion in 2016. In particular, entry of online classified-ad competitors such as Craigslist substantially reduced newspapers' revenue. Even if the share of digital advertising revenue has increased from 17% in 2011 to 29% in 2016 in the U.S., it is far from reversing the downfall in advertising revenue.[2] As a consequence, newspaper newsroom employment in the U.S. decreased by 37% for the period of 2004-2015 from 65,440 to 41,400.

Newspapers are in stiff competition with new online media. Among online media sources, news aggregators seem to be the most important. According to an Outsell report (2009), 57% of news media users go to digital sources, and they are also more likely to turn to an aggregator (31%) than to a newspaper site (8%) or other news site (18%). Indeed, Reuters Institute (2015) shows that aggregators (Yahoo! News, Google News, MSN, Buzzfeed and Huffington Post) attract 80% of the online news traffic in the U.S. In South Korea, the two major news aggregators, Naver and Daum, each had a share of 55.4% and 22.4% in the Internet news traffic in 2016 (Choi, 2017) whereas the traffic to newspaper home pages had only 4% share in 2017 (Korean Press Foundation, 2017).

The success of news aggregators has generated a heated debate about the effects of news aggregators on newspapers' incentive to produce high-quality content. During 2009 to 2010, the FTC hosted three workshops and published a controversial discussion draft (FTC, 2010) that hints at copyright reform and the protection of newspapers from aggregators. In Europe, the German Parliament introduced in 2013 a change in the copyright law that allowed news aggregators to link for free the news stories of news outlets if using excerpts of less than 7 words. Longer excerpts or images would require the payment of a negotiated fee to the news outlets. In 2014, a reform of the Spanish intellectual property law established that firms posting links and excerpts of news stories have to pay a compulsory link fee to the original publishers. In December 2014, Google reacted by shutting down Google News in Spain.

In the debate on news aggregators, content producers argue that news aggregators make money by stealing high-quality content. Since this money is

[2] According to *The Economist* ("Taxing Times", 10 Nov. 2012), in France, not a single national newspaper is profitable despite around € 1.2 billion in direct and indirect government subsidies.

pulled out of content producers' pockets, they have less incentive to produce high-quality content. For instance, according to Rupert Murdoch, chairman of News Corp.:

> "When this work is misappropriated without regard to the investment made, it destroys the economics of producing high-quality content. The truth is that the 'aggregators' need news organizations. Without content to transmit, all our flat-screen TVs, computers, cell phones, iPhones and blackberries, would be blank slates. (Murdoch, 2009: 13)."

On the other hand, news aggregators argue that aggregation drives profitable traffic to news sites themselves. In a response to the FTC report (2010), Google (2010) claimed to send more than four billion clicks per month to news publishers via Google Search, Google News, and other products. Google's claim is that each click –each visit– provides publishers with an opportunity to show ads, register users, charge for access to content, and so forth.

In this survey, I review empirical and theoretical articles on news aggregators. The empirical articles aim at quantifying whether news aggregators steal traffic from newspapers or help them to receive more traffic. In other words, they study which effect dominates between the two opposite effects, *i.e., the business-stealing effect and the readership-expansion effect which* I introduce in Section II. The theoretical articles aim at identifying different channels through which news aggregators affect profits of newspapers in order to analyze how news aggregators affect newspapers' incentive to invest in quality.

There exist a variety of news aggregators. Some, like Huffington Post, use editorial staff, while others, like Google News, use an algorithm to find high quality content. After finding high quality articles, each aggregator posts them on its site. This, however, can be done in different ways. Some, like Yahoo! News, post the whole article on their site, with no link to the original content. Usually, this is because the aggregator pays the newspaper for that content and hence has the right to publish it. In 2006, Yahoo! signed an agreement with Newspaper Consortium[3] to use their content. Others, like Google News, show the title and a short summary and provide a link to the original article. These two types of aggregators bring revenue to newspapers in different ways: the first by buying a content license, and the second by sending traffic to newspaper sites. This is why Yahoo! News has kept its service in Spain while Google News has been shutdown in Spain.

[3] http://www.npconsortium.com/ "Is Yahoo a Better Friend to Newspapers Than Google?", *New York Times,* 8 Apr. 2009.

The survey is organized as follows. In Section II, I present a simple theory, which provides a framework to understand the empirical findings reviewed in Section III. I review theoretical studies in Section IV. Section V provides concluding remarks with some questions for future research.

II. A SIMPLE THEORETICAL FRAMEWORK

Let me start by providing a simple theory based on Jeon and Nasr (2016) that captures the two opposite effects of news aggregators, the business-stealing effect and the readership-expansion effect. We consider two (major) newspapers and one aggregator and study their competition on the Internet. Suppose that the two newspapers compete on the Hotelling model. The two newspapers are located at the extreme points of a line of length one: newspaper 1 (2) on the left (right) extreme point. The line represents ideological differentiation (Mullainathan and Shleifer, 2005 and Gentzkow and Shapiro, 2011) and the insights would hold even if the two newspapers' locations are not extreme. A mass one of consumers are uniformly distributed over the line. We assume that consumers single-home, which means that without (with) the aggregator, a consumer consumes only one between the two newspapers (among the two newspapers and the aggregator).

We assume for expositional convenience that there is a continuum of topics which each newspaper covers. Let S be the set of topics. A topic can be about an election, an earthquake, a sport event, the climate change etc. On each given topic, a newspaper can provide either high or low quality content. So the strategy of newspaper i, with $i \in \{1, 2\}$, is a subset of topics $s_i \in S$ which it covers with high quality. Let $\mu(s)$ represent the measure of any set $s \in S$. Without loss of generality, assume $\mu(S) = 1$. Then, $\mu(s_i)$ represents the average quality of newspaper i. In addition to this vertical dimension of strategy, there is an horizontal dimension of strategy. Namely, given $0 < \mu(s_1), \mu(s_2) \leq 1/2$, for newspaper $i \in \{1, 2\}$, if i chooses s_i such that $s_i \cap s_j = \emptyset$, we say that i uses the *maximum differentiation strategy*. If i chooses s_i such that $\mu(s_1 \cap s_2) = \min(\mu(s_1), \mu(s_2))$, then we say that i uses *the minimum differentiation strategy.*

Let $u_0 > 0$ represent a consumer's utility from reading the home page (or the landing page) of a newspaper. u_0 is assumed to be large enough to make all consumers consume a newspaper or the aggregator. The home page provides links to articles with their titles and excerpts. Consumers are assumed to click a link only if the article is of high quality. Let $\mu\Delta u > 0$ represent the utility increase (net of attention cost) a consumer experiences from reading measure μ articles of high quality. If the quality of an article is low, no consumer reads it. Then,

the utility that a consumer located at x obtains from consuming newspaper 1 or 2 is given by

$$U^1(x) = u_0 + \mu(s_1)\Delta u - xt; \quad [1]$$

$$U^2(x) = u_0 + \mu(s_2)\Delta u - (1 - x)t, \quad [2]$$

where $t > 0$ is the transportation cost parameter and xt (or $(1 - x)t$) represents the cost of imperfect match in terms of ideological preferences.

In the absence of the aggregator, given ($\mu(s_1)$, $\mu(s_2)$), the market share of newspaper i is determined by

$$\alpha_i^N = \frac{1}{2} + \frac{\Delta u}{t}\left(\mu\left(s_i\right) - \mu\left(s_j\right)\right)$$

where the superscript N means no aggregator and $i, j = 1, 2$ and $i \neq j$.

We consider free newspapers which make revenue from advertising. The advertising revenue of a newspaper is assumed to be proportionate to the attention that consumers spend on the newspaper: but a given unit of attention spent on a home page may generate a larger (or lower) revenue than the same unit of attention spent on individual articles. The advertising revenue generated by a consumer's consumption of a home page is normalized to one. We assume that if a consumer consumes μ measure of high quality articles, it generates an advertising revenue of $\delta\mu$. Therefore, newspaper i's profit without the aggregator is given by

$$\pi_i^N = \alpha_i^N\left[1 + \mu\left(s_i\right)\delta\right] - c\left(\mu\left(s_i\right)\right), \quad [3]$$

where $c(\cdot)$ is the cost of producing high quality articles and is increasing and convex.

We model an aggregator along the lines of Google News in that the aggregator provides only a home page without having its own original articles. For each topic, the aggregator chooses one article and publishes its title and its excerpts (called also, snippets) with a link to the original article. We assume that the aggregator chooses the highest quality article for each topic and that if both newspapers produce the same quality articles on a given topic, it chooses one of them with an equal probability.

A consumer who reads the aggregator's home page obtains a utility of $u_0 + u_T$ where $u_T > 0$ is the utility from the aggregation of content from third-

parties, *i.e.,* numerous small news sites different from newspaper 1 and 2. In addition, she clicks on the link of each high quality article and spends attention on the newspaper site to which she is directed. The consumer is assumed not to click on the links to low quality articles. Therefore, using the aggregator over her preferred newspaper allows a consumer to access more high quality content, at a higher cost of preference mismatch.

More precisely, consider a consumer with location $x < 1/2$. Then, we have

$$U^{Agg}(x) - U^{1}(x) - u_T = \underbrace{\left(\mu(s_1 \cup s_2) - \mu(s_1)\right)\Delta u}_{\text{Benefit from higher quality}} - \underbrace{t\left(\frac{1}{2} - x\right)\left(1 + \mu(s_2) - \mu(s_1)\right)}_{\text{Cost from higher preference mismatch}}, \qquad [4]$$

where $U^{Agg}(x)$ represents the utility that a consumer located at x obtains from using the aggregator (see the appendix for the explicit formula). The benefit of using the aggregator instead of newspaper 1 is composed of u_T and the other terms. The term $(\mu(s_1 \cup s_2) - \mu(s_1))\triangle u$ represents surplus increase from consuming more high quality content. This benefit comes with the cost of greater preference mismatch since, for a consumer with location $x < 1/2$, the favorite newspaper is 1; the last term in [4] always has a negative sign for $x < 1/2$.

Jeon and Nasr (2016) assume that producing high-quality articles is costly such that each newspaper i chooses $\mu(s_i) \leq 1/2$. They show that under reasonable assumptions, the maximum differentiation strategy is a dominant strategy for each newspaper as this strategy allows each newspaper to maximize the traffic directed from the aggregator to its individual articles.

Under the maximum differentiation strategy, given (s_1, s_2), newspaper i's profit is given by:

$$\pi_i^A(s_i \mid \max) = \alpha_i^A\left[1 + \mu(s_i)\delta\right] + \delta\mu(s_i)\left(1 - \alpha_i^A - \alpha_j^A\right) - c\mu(s_i)^2 \qquad [5]$$

where $j \in \{1, 2\}$, $j \neq i$ and the superscript A means that the aggregator is present. When compared with the profit without the aggregator [3], the term in the middle of the R.H.S. of [5] is new and represents the advertising revenue from the consumers directed by the aggregator to i's articles as $\left(1 - \alpha_i^A - \alpha_j^A\right)$ represents the aggregator's share in home page traffic.

Both the *business-stealing effect* and the readership-expansion effect are defined in terms of traffic. The business-stealing effect captures the reduction in the traffic to the home pages of the newspapers which results as some consumers read the home page of the aggregator and is given by $\alpha_i^A - \alpha_i^N\left(= -\Delta\alpha_i\right) < 0$. *The readership-expansion effect* captures the traffic increase to high quality articles

which result as high quality articles of a newspaper can reach not only its loyal readers, but also those using the aggregator. The latter includes consumers who would read the rival newspaper if there were no news aggregator. If traffic is measured in terms of the number of pageviews, $\Delta\alpha_j\mu(s_i) > 0$ represents the readership-expansion effect to newspaper i where $\Delta\alpha_j$ represents the consumers who switch from newspaper j ($\neq i$) to the aggregator in terms of home page consumption. Note that in Jeon and Nasr (2016), the total number of consumers is fixed and hence the readership-expansion effect means that consumers on average read more articles. However, one can also consider another kind of readership-expansion effect, which means that the aggregator increases the number of consumers who read news (see Dellarocas, Katona and Rand, 2013 in Section IV).

The empirical papers reviewed in Section III try to study which of the two effects dominates. In addition, they try to see how the effects interact with the characteristics of newspapers. For instance, within the simple framework presented in this section, the small (unknown) newspapers whose content is aggregated by the aggregator and is captured by u^T for sure gain from the presence of the aggregator as they attract no traffic in its absence. Although this extreme result is obvious and is driven by assumption, I will provide some empirical evidence for a more generalized version of the result in the next section.

In the end, what matters for each newspaper is how its profit is affected. Given (s_1, s_2), the effect of the aggregator on newspaper i's profit is given by

$$\pi_i^A\left(s_i \mid \max\right) - \pi_i^N\left(s_i\right) = \underbrace{-\Delta\alpha_i}_{\text{Business-stealing effect (-)}} + \underbrace{\delta\Delta\alpha_j\mu\left(s_i\right)}_{\text{Readership-expansion effect (+)}}$$

where $\delta > 0$ captures the monetary value of a unit traffic to articles relative to that of a unit traffic to home page in terms of advertising revenue and is typically smaller than one. Hence, even if the total effect on the traffic of newspaper i is positive, the total effect on its profit can be negative.

The theoretical papers reviewed in Section IV investigate how the aggregator affects each newspaper's incentive to invest in quality, which can be studied only after one understands how the aggregator affects each newspaper's profit for given quality choices. However, as it is hard to find data on profits, the empirical papers seldom study the effect on profits. Therefore, there is a gap between theoretical papers and empirical papers. This is why I review first the empirical papers before reviewing theoretical papers.

Finally, each newspaper can employ another strategy, which consists in opting out from the news aggregator. Jeon and Nasr (2016) show that if an increase in the third-party content indexed by the aggregator generates more traffic to each newspaper, then each newspaper has no incentive to opt out. Opting out implies losing traffic from the aggregator. This adverse effect of the opting out should increase with the market share of the aggregator, which in turn increases with the amount of the third-party content indexed by the aggregator (represented by u_T). I will review below an empirical study of opting in/out decisions in Germany by Calzada and Gil (2017).

In summary, the simple theoretical framework generates the following questions to be answered. On the empirical side, we have:

- Which effect dominates between the business-stealing effect and the readership-expansion effect?
- How do the effects vary depending on the characteristics of newspapers?
- Does a newspaper have an incentive to opt-out?

On the theoretical side, we have:

- How does the aggregator affect each newspaper's profit?
- How does the aggregator affect each newspaper's incentive to invest in quality?

III. EMPIRICAL STUDIES OF NEWS AGGREGATORS

In this section, I review empirical studies of news aggregators. I start by reviewing papers that study Google News: Google News shutdown in Spain, Google News opt-in policy in Germany and other events related to Google News. And then, I review a paper that studies Facebook as a news aggregator and an experimental paper studying attention allocation between a news aggregator and original articles. Finally, I review a paper studying news slants of aggregators.

Before reviewing the empirical results, let me point out the fact that all empirical papers find that the business-stealing effect is dominated by the readership-expansion effect.

1. Empirical Studies of Google News

1.1. Events Regarding Google News in Spain and in Germany[4]

Let me first describe the events regarding Google News in Spain and in Germany. On January 1, 2014, because of the lobbying of the publishers' association AEDE, the Spanish Parliament passed a reform of the law of intellectual property right. The new law established that online outlets posting links and excerpts of news articles originated elsewhere must pay a link fee to the original publishers. A unique feature of the Spanish regulation is that link fees are mandatory: publishers cannot refuse to receive a fee from news aggregators as the link fee must be collected by a private entity called CEDRO which will redistribute the revenues to the news outlets. (Calzada and Gil, 2017).

Although the implementation of the law was subject to a lot of uncertainty, on December 16, 2014, Google shut down the Spanish edition of Google News. The shutdown had an important and immediate impact on the Spanish news market such that the publishers in AEDE urged the government to negotiate a solution with Google. Some large publishers in AEDE even announced that they would renounce any compensation payment for sharing content with news aggregators.

The German Parliament passed an addendum to the copyright law on March 1, 2013. It granted publishers the right to charge search engines and other online aggregators for reproducing their content beyond headlines and short excerpts but also allowed free use of text in links and brief excerpts. The main differences of the German regulation with respect to the Spanish one are: link fees have to be negotiated between the parties and brief excerpts are not affected by the regulation.

In June 2014, VG Media, a consortium of more than 200 publishers, sued Google and other news aggregators for displaying excerpts and preview images along with the links to their news articles. On October 2, 2014, the German edition of Google News announced the change from an opt-out to an opt-in system: those publishers who want to be indexed by Google must explicitly grant permission and renounce any type of compensation. Publishers associated with VG Media decided not to opt in. A leading publisher in the group of VG Media was Axel Springer, which asked VG Media not to issue free licenses for its websites. On October 23, 2014, Google News and other German news aggregators stopped showing large excerpts, video and images from the

[4] The description of the events is mainly based on Calzada and Gil (2017).

publishers that did not opt in. The change significantly reduced traffic to VG Media news sites that on November 5, 2014, Axel Springer and other VG Media publishers decided to opt in.

1.2. Google News shutdown in Spain

Athey, Mobius and Pal (2017) study Google News shutdown in Spain by using browser log data of desktop users. Control users are chosen to have identical news consumption patterns as treatment users after the shutdown. Before the shutdown, treatment users used Google News whereas control users did not. They estimate the effect of the shutdown by comparing the news consumption of treatment and control users before the shutdown.

They find that treatment users have 19.7% higher consumption in terms of pageviews in the pre-shutdown period compared to control users, including their consumption of the Google News home page. This volume change comes from two sources: Google News users consume 28.8% more articles but 8.5% fewer landing pages (omitting the Google News landing page). Hence, the readership-expansion effect dominates the business-stealing effect: in other words, Google News is a complement to overall news reading.

Athey, Mobius and Pal (2017) also break out the volume effect by distinguishing top 20 outlets from below top 20 outlets. They find that the effect of Google News on the top 20 outlets is not statistically different from zero as the positive effect on articles cancels out the negative effect on landing pages. By contrast, smaller outlets gain as much as 26.3% from the presence of Google News: the landing page traffic is unaffected but article pageviews increase by 44.6%. They further decompose the volume effect according to news characteristics. They find that post-shutdown, treatment users read less breaking news, hard news and news that is not well covered on their favorite news publishers.

Calzada and Gil (2017) use data at the domain level from news outlets in Spain, France and Germany. Hence, their data are complementary to the data used by Athey, Mobius and Pal (2017). They study the Google News shutdown in Spain by using French outlets as a control group. They find that the shutdown reduced on average the number of daily visits to Spanish outlets by 14%. This finding is consistent with that of Athey, Mobius and Pal (2017). Calzada and Gil (2017) find that this effect varies from no effect (business outlets), medium size effect (national and regional news outlets) and large effect (sports and Catalan language news outlets). They also find that the impact was larger in lower-

ranked domains and domains with lower proportion of international visitors, which is quite consistent with the finding of Athey, Mobius and Pal (2017).

Calzada and Gil (2017) also study how the impact of the shutdown evolved over time until reaching a steady state. They find that the effect across all news outlets stabilizes around 13.8% seven weeks after the shutdown. They also try to decompose the total effect into a market-expansion effect and a substitution effect by studying the impact of the shutdown on the outlets' traffic sources. They find that the percentage of search visits decreased whereas the percentage of direct visits increased. They interpret the former as an evidence of the market-expansion effect and the latter as an evidence of the substitution effect.

Whereas Athey, Mobius and Pal (2017) limit attention to the impact of the shutdown on traffic, Calzada and Gil (2017) study the impact on advertisement revenues as well. They focus on the online editions and separate those outlets that are above the median advertising revenues from those that are below the median. They find that after the shutdown, the daily revenues of above median outlets decreased significantly relative to those below the median. When they study the sources of this decrease in revenue, they find decreases in advertising intensity, revenue per advertiser and revenue per unit of advertising intensity. However, it seems that their finding on advertising revenue is hard to reconcile with the finding of Athey, Mobius and Pal (2017) that the shutdown did not change the overall traffic but increased the traffic to landing pages for top 20 outlets whereas it reduced the overall traffic without affecting the traffic to landing pages for below top 20 outlets. Suppose that top 20 outlets have advertising revenues above the median and that landing pages are more important than individual articles in terms of advertising revenue. Then, the shutdown should increase advertising revenue for top 20 outlets while reducing it for below top 20 outlets, which is opposite to the finding of Calzada and Gil. It would be nice to have a better understanding of the impact on advertising revenue.

1.3. Google News opt-in Policy in Germany

After the introduction of the opt-in policy in Germany, Google News continued to index all news outlets but could complement the links with long excerpts and images only from those outlets that had opted in. Calzada and Gil (2017) study the impact of VG Media's decision not to opt in. They find a negative but non-significant effect of the opt-out decision on the visits to the VG Media outlets relative to all other German outlets that did not belong to VG Media. But when they focus on the 10 outlets Axel Springer controlled, which are part of the VG media outlets, they find a negative and significant reduction in daily visits of around 8% in Axel Springer outlets relative to all other German

outlets. This explains the fact that Axel Springer and the other VG Media outlets that had initially stayed out decided to opt in. The scenario that some outlets opt in while others do not is hard to be sustained as an equilibrium as the latter has competitive disadvantage because the traffic they would receive from Google News with opt-in is likely to be directed to the former. What happened in Germany is consistent with the prediction of Jeon and Nasr (2016).

1.4. Other Studies on Google News

Chiou and Tucker (2017) study the removal of the content of Associated Press (AP) from Google News that occurred from December 23, 2009 until sometime in February 2010. They use Yahoo! News as a control since it continued to host the AP content. They study whether the removal leads to a shift away from Google News and whether traffic to news sites from Google News falls after the removal. They find that the removal does not affect the traffic to Google News. In the case of the effect on downstream news sites, they find that the odds of visiting a news site on Google News relative to a non-news site on Google News decreased by 28% compared to the odds of visiting a news site on Yahoo! News relative to a non-news site on Yahoo! News. This result suggests that the presence of AP articles in Google News prompted users to seek further information at news sites.

One striking feature of how AP content was featured on Google News is that in general quite a large amount of news content was displayed rather than merely a snippet. In light of this, the result that Google News increases traffic to downstream news sites is surprising. It is even more surprising in view of the finding of Dellarocas *et al.* (2016) that a longer snippet reduces the probability of clicking on the link (see Section III.3).

Athey and Mobius (2012) study a case where Google News added local content to its home page for those users who chose to enter their location. By comparing the consumers who use this feature with controlled users, they find that users who adopted the feature increased their usage of Google News, which in turn led to additional consumption of local news. They conclude that their results support the view that news aggregators are complementary to local news outlets.

George and Hogendorn (2013) use a major redesign of Google News on June 30, 2010 that placed a permanent strip of geo-targeted local news headlines and links onto the Google News front page and find that adding geo-targeted links increases both the level and share of local news consumed online.

2. Facebook as a News Aggregator

Sismeiro and Mahmood (2018) study how an outage of Facebook affected traffic to a news website. They have traffic data from the second largest online news website operating in a major Western European country. They take advantage of the exogenous variation in Facebook traffic created by a global Facebook outage that lasted four hours in the early morning of Monday, October 21, 2013. During the outage, it was not possible to add new posts, comment on previous posts and there were no newsfeed updates although users could access the information previously loaded on their device. Their data cover the period of October 13, 2013 to October 29, 2013 (17 days).

They observe a 38% decrease in visitors per hour and a 44% reduction in the total number of page views during the outage and a drop of about 9% of page views even after the outage. The results suggest that Facebook helps news websites to attract visitors and leads to more page requests. More importantly, they find that Facebook has an effect that goes beyond the traffic originating from clicks on the links to the news site posted on Facebook. This is because an hourly decrease of 3,956 page views originates directly from Facebook during the outage, which is substantially lower than the reduction in total page views during the outage (about 170,000 pages). More precisely, they find that during the outage hours, referrals from search engines and undefined referrals (*i.e.,* people directly typing the URL, using their own bookmarks, or copying and pasting URLs) decreased far more than Facebook referrals. They find 29,470 fewer referrals from search and 142,020 fewer undefined referrals. This seems to be an interesting finding which shows a main difference between Facebook and Google News in terms of how each affects traffic to news sites.

However, the result may be due to the so-called "dark traffic" problem, which arises when a huge proportion of referral traffic is listed as "direct". Research from the analytics firm Chartbeat, as well as confirmation from major publishers, shows that Facebook's mobile apps are largely responsible for the swathes of dark traffic being directed toward websites.[5] Hence, most undefined referrals are likely to be originated from Facebook.

They further look at the performance of different news categories during the outage. The news categories they study include local news, sports, women issues and health. They find a reduction in traffic of all news categories during the outage. In contrast, after the outage, traffic recovery varies by category. Sports and local news see a significant increment after the outage whereas women issues and health-related sections remain below the baseline. They

[5] See http://uk.businessinsider.com/facebook-mobile-app-responsible-for-dark-traffic-2014-12?r=US&IR=T

speculate that this difference arises because the first two categories are more time sensitive than the last two.

They also find that during the outage, a decrease in the number of home page views per user of 0.71 and an increase in the number of content page views per user of 0.52. These correspond to a reduction of 66% and an increase of 37% compared to their baselines. This suggests that Facebook introduce a selectivity bias by attracting shallower users (*i.e.,* users who read mostly headlines from the home page and do not read many articles) to the site.

I think that the result that the Facebook outage reduced the traffic to the news site is much less surprising than the findings from the Google News shutdown in Spain as the former is about a temporary shock while the latter is about a lasting or permanent shock.

3. Experiments on Attention Allocation Between a News Aggregator and Original Articles

Dellarocas *et al.* (2016) study how readers allocate their attention between a news aggregator and the original articles it links to. They run field experiments on a Swiss news aggregator application called Newscron. The app has two client versions, an iPhone version and an iPad version. The two versions provide distinct user interfaces with different limitations and hence they conducted separate experiments on each version.

They first consider topics that have a single article in the iPhone environment and find that click-through probabilities of individual articles decrease as snippet lengths increase and that the presence of an image is also associated with lower click-through rates. Experiments with the iPad version lead to the same results. These findings suggest that click-through rates are significantly affected by snippet lengths. However, one can expect that the snippet length which is optimal for the newspapers providing original articles is shorter than the one which is optimal for the aggregator. This may provide a rationale for regulating the snippet length as is done in Germany.

They also consider topics containing two or more snippets and where exactly one snippet was clicked and study how an article's snippet length and the presence of an image affect the click-through probability in the iPhone environment. As only one snippet in a topic is clicked, this study allows them to study how competition among snippets is affected by snippet length and image. They find that having longer than average snippets has a positive effect

on the choice probability and that the presence of an accompanying image increases a snippet's within group choice probability. The effect of having an image is strong and comparable to moving from second to first position on the list of related articles.

This result on click-through probability in a competitive environment is consistent with the finding of Calzada and Gil (2017) that those newspapers which opted out (and hence whose articles had very short snippets on Google News) suffered from traffic loss. Because of this competitive disadvantage, they ended up opting in. The same analogy can be made to the Swiss news aggregator in Dellarocas *et al.* (2016): even if newspapers may collectively prefer short snippets, short snippets may not be sustained as an equilibrium when each news site can deviate by allowing the news aggregator to show longer snippet.

4. News Slant of Two Korean Aggregators

South Korea is unique in terms of the influence of news aggregators. In 2016, 60% of Koreans had access to news through Internet portal news aggregators while only 13% consumed news through home pages of newspapers. The two major news aggregators, Naver and Daum, each had a share of 55.4% and 22.4% in the Internet news traffic in 2016 (Choi, 2017). The business model of Naver and Daum is similar to that of Yahoo News in that each of them pays to receive articles from a selected group of newspapers. In 2015, 59 newspapers supplied articles to both aggregators while 17 only to one aggregator and 86 (60 among them sports or entertainment newspapers) only to the other aggregator.

Choi (2017) studies news slants of the two Korean news aggregators by adopting the methodology of Gentzkow and Shapiro (2010). He has data about all news articles shown by both aggregators during 2015 and he finds that both of them exhibit almost no slant. Even if there is competition between the two Korean news aggregators, Choi (2017) finds little ideological difference between the two. This finding is very consistent with the theoretical prediction of Gabszewicz, Laussel and Sonnac (2001) that when newspapers are financed by advertising, they tend to have minimal ideological differentiation instead of the maximal differentiation, which occurs when they are financed by sales revenue (Mullainathan and Shleifer, 2005). A main difference between Choi (2017) and Gabszewicz, Laussel and Sonnac (2001) is that in Choi (2017), the slant of an aggregator is defined as the average slants of all articles shown by the aggregator, which are supplied by different newspapers which can have very strong ideological bias.

IV. THEORETICAL STUDIES

Most theoretical articles on news aggregators go beyond the empirical articles surveyed in Section III in the sense that they are not only interested in identifying different channels through which aggregators affect traffics and profits of newspapers but also interested in studying how the aggregators affect quality choices of newspapers, which is a very important question. Note also that most theoretical articles reviewed in this section consider a single-topic model whereas Jeon and Nasr (2016) consider a multi-topic model. How the results obtained in a single-topic model can be generalized to a multi-topic environment remains an open question.

In the model presented in Section II, Jeon and Nasr (2016) study how the aggregator affects the newspapers' incentive to invest in quality. They find that depending on the value of δ, it can increase or decrease the quality since the readership-expansion effect becomes stronger as δ increases. In order to further pin down the prediction, they find a lower bound on δ from the empirical findings of Athey and Mobius (2012) and Chiou and Tucker (2017). For instance, Athey and Mobius (2012) find that after adding content from new local outlets to Google News, traffic increases not only to these new outlets but also to the old (local and non-local) outlets that have been indexed by Google News. Using the lower bound, they find that the aggregator increases the quality chosen by each newspaper. They also find that the result on quality choice is robust to introducing noise into the quality certification technology of the aggregator. However, noise in the certification technology makes the business-stealing effect stronger relative to the readership-expansion effect, which tends to decrease newspapers' profits. This finding offers a possible explanation for newspapers' complaint against Google News: they may find Google's algorithm to select news articles too noisy, resulting in low profits for them.

Huang (2017) focuses on how a news aggregator alleviates the moral hazard of a newspaper in terms of investment in quality. In Jeon and Nasr (2016), the quality of each newspaper is known to consumers before they choose a newspaper to read. In her model, consumers do not observe the quality of a newspaper when they decide to visit its site or not. Hence, in the absence of the aggregator, the market collapses as the newspaper cannot commit to invest in quality: shirking is a dominant strategy for the newspaper. The aggregator alleviates this incentive problem as consumers can observe the quality of a newspaper by visiting the aggregator and can click on the link only if the quality is high. Depending on the degree of loyalty to the newspaper, a consumer can directly visit the newspaper or visit the newspaper indirectly by clicking the link at the site of the aggregator or visit only the aggregator. In addition, motivated by Dellarocas *et al.* (2016), she allows the aggregator to choose the length of

snippet. She finds that the aggregator tends to choose a snippet length which is too long as it does not internalize the traffic directed to the newspaper. Therefore, it can be optimal to introduce a tax on the snippet length or a click-through subsidy. The information asymmetry problem she studies should be relevant for those newspapers with weak brand recognition. Then, the aggregator can help consumers discover interesting articles from these newspapers as is found by Athey, Mobius and Pal (2017). Regarding snippet length, it would be interesting to empirically study whether the snippet length chosen by aggregators is too long. Note that an aggregator has some incentive to limit snippet length of the articles shown at its home page for the same reasons as all newspapers limit snippet length of the articles shown at their home pages. The two major Korean news aggregators, Naver and Daum, are an extreme example since they show only one line for each article in the mobile home page. In fact, they do not even show the source of each article.[6]

While Jeon and Nasr (2016) consider homogenous consumers (but for their ideological taste), Rutt (2011) considers two types of consumers (loyal ones and searchers) and uses an all-pay auction model to study newspapers' choice of quality and prices. A loyal consumer reads only her preferred newspaper while a searcher uses an aggregator to read the highest quality one among free newspapers as searchers are assumed to be not willing to pay to access an article. Given the behavior of consumers, firms simultaneously decide on their price and quality investments. Firms face a trade-off in their pricing strategy between earning sales revenue from loyal consumers and losing potential advertising revenue from searchers, which leads to a symmetric mixed strategy equilibrium. In the equilibrium, firms randomize between providing the article for free and charging for access to the article. There is a unique level of quality provided by the firms who charge for access to the article whereas there is a distribution of quality levels for articles which are free to access. He finds that as the fraction of searchers increases, the expected profit of each newspaper decreases, free newspapers choose higher quality while the rest choose lower quality. Although the results are interesting, I wonder how realistic the mixed strategy result is. The decisions regarding business models (free or paywall) and quality investments are core long-term decisions of a newspaper. For instance, the quality investment decision is strongly associated with the number of journalists to hire. I have difficulty in imagining a board taking these decisions in a random way.

Dellarocas, Katona and Rand (2013) go beyond a standard model of media and aggregators by considering competition among content sites in a link

[6] For some major topics, clicking on a topic at the homepage opens a second page about the topic showing multiple articles. Even in this case, they use snippets of only one or two lines per article (but provide the sources of articles as well).

economy. They consider a single-topic model. Each content site i can produce its own content of quality q_i and also provide a link to the content of another content site say j. Content site j cannot refuse the link of i. Content site i faces the following trade-off when providing a link to content of higher quality: it increases the anchor traffic to site i but a fraction $1 - \rho \in (0, 1)$ of the traffic will click the link and hence will not stay at site i meaning that site i obtains no advertising revenue from that traffic. After studying the equilibrium quality choice and link decision without aggregator, they introduce an aggregator who is defined as a content site which cannot create its own content but can provide a link. The aggregator provides the link to the highest quality content. By so doing, it increases the total anchor traffic to the media ecosystem (*i.e.,* there is a market-expansion effect) but reduces the anchor traffic to each content site. In addition, a fraction $1 - \rho$ of the anchor traffic of the aggregator ends up landing at the highest quality site. If $1 - \rho$ is large enough, the aggregator increases the traffic to the highest quality site while always reducing the traffic to the lowest quality site. When the content sites cannot provide links (like most newspaper sites in real world), they find that the equilibrium content quality decreases with ρ. They also consider imperfect quality certification technology of the aggregator and find that as the technology becomes more accurate, there is more competition between the content sites such that the equilibrium quality becomes higher and the profit becomes lower. This result is opposite to the finding of Jeon and Nasr (2016). It would be interesting to dig deeper into the role of the algorithm used by the aggregator.

De Cornière and Sarvary (2018) study content bundling by social media, *i.e.,* social media shows news content together with user-generated content (UGC). In the baseline model, they consider one newspaper. They are interested in studying how the content bundling affects the profit of the newspaper and its incentive to invest in quality. UGC quality is assumed to be exogenous. Each consumer allocates a fixed total amount of attention between news and UGC and consumers differ in terms of their demand intensity for news. In the benchmark without content bundling, consumers optimally allocate their time between social media to consume UGC and the newspaper site to consume news. In order to understand the effect of content bundling, we can consider personalized content bundling: the social media knows each consumer's type and bundles a different amount of news content depending on the type. In this case, it is optimal for the social media to propose exactly the same amount of news content each type will consume in the benchmark without content bundling. This reduces the newspaper's profit because for any news consumed on the social media, the associated advertising revenue is shared with the social media. Even though this effect tends to reduce the incentive for the newspaper to invest in quality, however, there is an opposite effect which makes the overall effect on the investment incentive ambiguous. Namely, by investing more,

the newspaper can induce those consumers who spend very small amount of attention on UGC to spend their entire attention directly on the site of the newspaper. This increases the advertising revenue of the newspaper in a discrete way as the newspaper captures all advertising revenue associated with news consumption on its site. They find qualitatively the same results on the profit and the investment incentive when the social media cannot personalize content bundling.

Calzada and Ordóñez (2012) study a newspaper's reaction to the aggregator in terms of versioning (and linking) decisions in the framework of a monopolist's second-degree price discrimination. George and Hogendorn (2012) consider a model of two-sided market in which news aggregators increase multi-homing viewers. They find that the switching of a given mass of viewers from single-homing to multi-homing is likely to reduce (increase) a news outlet's advertising revenue if the outlet initially has a high (small) share of exclusive viewers.

V. CONCLUDING REMARKS

A big challenge for newspapers in the Internet environment is how they can attract attention of consumers who spend their limited attention among millions of different sites. For instance, Boik, Greenstein, and Prince (2017) find that for the period of 2008-2013, total time online at the primary home device has only modestly declined and that the concentration of sites visited and time spent in long sessions has remained remarkably stable. Their finding implies that the total amount of attention that consumers spend on the Internet is more or less fixed and is concentrated on a relatively small number of anchor sites. This puts newspapers in a vulnerable situation as they become dependent on major anchor sites such as Facebook and Google (Search and News) to attract traffic to their news sites. Such trend is observed by Boik, Greenstein, and Prince (2017) as they find that the period between 2008 and 2013 saw major changes in online category shares, with social media and video experiencing significant increases while chat and news experienced significant declines.

A major empirical finding I surveyed is that news aggregators reduce traffic to newspaper home pages while increasing traffic to individual news articles. Even if all empirical articles agree on the statement that the business-stealing effect is dominated by the readership-expansion effect, if this comes with a reduced traffic to home pages, it can have a long-term consequence that is not captured by the empirical studies. For instance, if consumers using news aggregators do not pay much attention to the sources of original articles, this can reduce newspapers' incentives to build up reputation, which would make newspapers further depend on the reputation of the aggregators such as Google

or Facebook. It would be interesting to study both empirically and theoretically how competition for attention among newspapers is done on a anchor site (such as news aggregators or social media) and how such competition is different from the competition among printed newspapers before the Internet. Is the competition on a news aggregator healthier than the competition among printed newspapers? One can further study how competition on a anchor site is affected by the site's algorithm (such as Facebook's newsfeed algorithm) and how the profit-maximizing algorithm differs from the welfare-maximizing one. For instance, Facebook recently announced an algorithm change which will de-prioritize videos, photos, and posts shared by businesses and media outlets in favor of content produced by a user's friends and family.

Google and Facebook launched respectively AMP (accelerated mobile page) project in 2016 and Instant Articles project in 2015. AMP and Instant articles host articles respectively on Google and Facebook for fast-loading of news in the mobile environment. It seems that Google's AMP project has received much wider support from publishers than Facebook's Instant Articles.[7] In fact, more than half of Facebook's launch partners on Instant Articles, including major newspapers such as *New York Times* and *Washington Post,* appear to have abandoned the format (Brown, 2018). The different outcomes may have to do with different business models embraced by Google and Facebook; while Google is attached to open web, Facebook is a closed system with the goal of getting people to spend more time inside its app in order to show more ads. However, even with the success of the AMP project, there are concerns about increasing dependence of media companies on the major platforms through "mediated advertising arrangements with accidentally enormous middlemen apps that have no special interest in publishing beyond value extraction through advertising (Herrman, 2015)".

Note that the decrease in traffic to newspaper home pages relative to traffic to individual news articles is a more general phenomenon, which is called the unbundling of journalism:

"It is a world of fragments, filtered by code and delivered on demand. For news organizations, said Cory Haik, senior editor for digital news at *The Washington Post,* the shift represents "the great unbundling" of journalism. Just as the music industry has moved largely from selling albums to songs bought instantly online, publishers are increasingly reaching readers through individual pieces rather than complete editions of newspapers or magazines."[8]

[7] https://digiday.com/media/how-google-amp-won-over-facebook/

[8] *New York Times,* "How Facebook Is Changing the Way Its Users Consume Journalism" by Ravi Somaiya, October 26, 2014.

How will the unbundling of journalism affect the incentive to produce high quality journalism? A popular view is that the traditional way of selling bundle of news developed a cross-subsidy system which allowed to finance costly investigative journalism. For instance, according to a report prepared for FCC,

> "A cross-subsidy system had developed: a consumer who bought the newspaper for the box scores was helping to pay the salary of the city hall reporter. Today, a reader can get a mobile app that provides only box scores (with second-by-second updates!). The bundle is broken– and so is the cross-subsidy. (Waldman *et al.,* 2011: 13)."

Does the end of the cross-subsidy system imply the end of investigative journalism?

BIBLIOGRAPHY

Athey, S., and M. Mobius (2012), "The Impact of News Aggregators on Internet News Consumption: The Case of Localization," *Working Paper.*

Athey, S.; Mobius, M., and J. Pal (2017), "The Impact of Aggregators on Internet News Consumption," *Working Paper.*

Boik, A.; Greenstein, S., and J. Prince (2017), "The Empirical Economics of Online Attention," *Working Paper.*

Brown, P. (2018), "More Than Half of Facebook Instant Articles Partners May Have Abandoned It," February 2, *Columbia Journalism Review*.

Calzada, J., and R. Gil (2017), "What Do News Aggregators Do? Evidence from Google News in Spain and in Germany," *Working Paper.*

Calzada, J., and G. Ordóñez (2012), "Competition in the News Industry: Fighting Aggregators with Versions and Links," *Working Paper,* 12-22, NET Institute.

Choi, D. O. (2017), "Internet Portal Competition and Economic Incentives to Tailor News Slant," *Korean Journal of Industrial Organization,* 25(2): 1-40 (in Korean).

Chiou, L., and C. Tucker (2017), "Content Aggregation by Platforms: The Case of the News Media," *Journal of Economics and Management Strategy,* 26: 782-805

De Cornière, A., and M. Sarvary (2018), "Social Media and News: Attention Capture via Content Bundling," mimeo.

Dellarocas, C.; Katona, Z., and W. Rand (2013): "Media, Aggregators, and the Link Economy: Strategic Hyperlink Formation in Content Networks," *Management Science,* 59(10): 2360–2379.

Dellarocas, C.; Sutanto, J.; Calin, M., and E. Palme (2016), "Attention Allocation in Information-Rich Environments: The Case of News Aggregators," *Management Science,* 62(9): 2543-2562.

FTC (2010), "Potential Policy Recommendations To Support The Reinvention Of Journalism," *Discussion Draft,* Federal Trade Commission.

Gabszewicz, J. J.; Laussel, D., and N. Sonnac (2001), "Press Advertising and the Ascent of the 'Pensée Unique'," *European Economic Review,* 45(4-6): 641–651.

Gentzkow, M., and J. M. Shapiro (2010), "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *Econometrica,* 78(1): 35-71.

— (2011), "Ideological Segregation Online and Offline," *The Quarterly Journal of Economics,* 126(4): 1799–1839.

George, L., and C. Hogendorn (2012), "Aggregators, Search and the Economics of New Media Institutions," *Information Economics and Policy,* 24(1): 40–51.

— (2013), "Local News Online: Aggregators, Geo-Targeting and the Market for Local News," *Working Paper*.

Google (2010), "Comments on Federal Trade Commission's News Media Workshop and Staff Discussion Draft on "Potential Policy Recommendations to Support the Reinvention of Journalism"," *Discussion Paper.*

Herrman, J. (2015), "The Next Internet is TV," February 5, the AWL, https://www.theawl.com/2015/02/the-next-internet-is-tv/

Huang, J. (2017),"Should Google Profit Like a Taxi Driver?," *Working Paper.*

Jeon, D.-S., and N. Nasr (2016), "News Aggregators and Competition Among Newspapers on the Internet," *American Economic Journal: Microeconomics,* 8(4): 91-114 .

Korea Press Foundation (2017), *Digital News Report 2017 South Korea.*

Mullainathan, S., and A. Shleifer (2005), "The Market for News," *American Economic Review,* 95(4): 1031–1053.

Murdoch, R. (2009), "From Town Crier to Bloggers: How Will Journalism Survive the Internet Age?," Speech Before the Federal Trade Commission's Workshop.

Outsell (2009), "News Users," *Discussion Paper.*

Pew Research Center (2017), State of News Media, www.pewresearch.org

Reuters Institute (2015), "Digital News Report," *Discussion Paper,* University of Oxford.

Rutt, J. (2011), "Aggregators and the News Industry: Charging for Access to Content," *Working Paper,* 11-19, NET Institute.

Sismeiro C., and A. Mahmood (2018), "Competitive vs. Complementary Effects in Online Social Networks and News Consumption: A Natural Experiment," *Management Science,* Articles Published in Advance.

Waldman S., and the Working Group on Information Needs of Communities (2011), *The Information Needs of Communities: The Changing Media Landscape in a Broadband Age,* Federal Communications Commission.

APPENDIX

Given (s_1, s_2), the utility that a consumer with location x obtains from using the aggregator is given by:

$$U^{Agg}(x) = u_0 + u_T + \mu(s_1 \cup s_2)\Delta u - \left(\mu(s_1 - s_2) + \frac{1}{2}\left[\mu(s_1 \cap s_2) + (1 - \mu(s_1 \cup s_2))\right]\right)xt - \left(\mu(s_1 - s_2) + \frac{1}{2}\left[\mu(s_1 \cap s_2) + (1 - \mu(s_1 \cup s_2))\right]\right)(1 - x)t,$$

where $s_1 - s_2$ means $s_1 \cap s_2^c$. $u_0 + u_T + \mu(s_1 \cup s_2)\Delta u$ represents utility from reading gross of the transportation cost. The transportation cost depends on the composition of the articles covered by the aggregator, and is equal to the measure of articles from newspaper 1 multiplied by xt plus the measure of articles from 2 multiplied by $(1 - x)t$.

PART IV

New Technologies

MACHINE LEARNING FOR ECONOMICS AND POLICY

Stephen HANSEN

Abstract

This chapter focuses on applications of machine learning algorithms for economic research and policymaking. It first introduces basic concepts in machine learning, whose main branches include supervised and unsupervised learning. The second half of the chapter discusses use cases and applications of machine learning algorithms. First, it discusses the quantification of unstructured data and how to recover information in a way that is useful for economists. The second application concerns new possibilities for measurement, where the combination of machine learning and new digital data, provides the opportunity to develop measures of objects like inflation and economic activity. The last two applications are related to forecasting and causal inference. The overall message of the chapter is that machine learning provides the tools needed to fully exploit the possibilities of rich new digital data sources.

Key words: Machine learning, digital data.

JEL classification: C55.

I. INTRODUCTION

Recent times have seen an astonishing growth in the production of data. More data was created in 2014 and 2015 than in the entire history of humankind beforehand, and by 2020 there will be approximately 44 zettabytes, or 44 trillion gigabytes, of data (Marr, 2015). Much of this explosion is due to digitization, as new technologies allow previously ephemeral human activities to be recorded. Messages and photos are now routinely sent via email or social media, which in turn allows them to be stored on servers indefinitely. Digital data of more direct economic relevance is also now increasingly available. Information from, for example, individual consumers' purchases, detailed product price histories, and rich administrative records is already beginning to transform empirical research in economics.

Along with the growth of data has come new empirical methods for analyzing it. The field of machine learning has developed rapidly in the past ten years in response to the digitization of data, and contributes many ideas to artificial intelligence, which is currently receiving much public attention. The relevance of these advancements for empirical research in economics is less clear. The bulk of machine learning methods have been developed by computer scientists, statisticians, and engineers, who typically have different goals than economists in conducting empirical work. This raises the question of what potential uses there are of machine learning in economics given its emphasis on causal inference and counterfactual prediction.

The first goal of this chapter is to introduce basics concepts in machine learning, and its first part focusses on this. The second goal is to reflect on its potential impact on economic research and public policy, and it does so through the discussion of several application areas. The discussion is non-technical and focused on broad ideas.[1]

There are several takeaway points. First, one important but sometimes underappreciated use of machine learning is the ability to use entirely new types of data. Modern econometrics typically uses data that is "regular": it can be represented in rectangular form with rows corresponding to individual observations and columns to variables. Moreover, variables are typically recorded as single, quantitative measurements like the total expenditure of households or the wages of employees. However, many of the newly available digital data sources do not have this format: text, satellite images, and web search profiles contain vast amounts of economically relevant information but

[1] Readers interested in a more technical, academic discussion can refer to several recent excellent surveys in the economics literature (for example, Einav and Levin, 2014; Varian, 2014; Mullainathan and Spiess, 2017).

have non-standard data structures. Machine learning can be used to extract the important information from these sources, and clean them for econometric analysis. The chapter illustrates several cases in which off-the-shelf approaches have been used to effectively do this.

Second, it is important to recognize that many machine learning methods are often not appropriate for the kinds of problems that economists confront. The chapter provides examples of this in forecasting and causal inference.

Third, despite the differing goals of machine learning and economics, specific ideas from machine learning can nevertheless be incorporated and extended to meet the needs of economic research. This process is only just beginning in economics, but is likely to hold the key for allowing economists and policymakers to fully exploit the potential of digital data.

II. WHAT IS MACHINE LEARNING?

There appears to be no single, agreed-upon definition of machine learning. A generic definition is the study of algorithms that allow machines to improve their performance in some given task as new data arrives. A more expansive definition from a popular textbook is that machine learning is "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" (Murphy, 2012). However, these definitions do not fully convey the differences between machine learning and econometrics. After all, the ordinary least squares regression model familiar to any undergraduate student in economics detects patterns in data, and has higher-quality estimates when estimated on larger datasets.

One area of difference between machine learning and econometrics is the role of statistical inference. Econometricians tend to focus on formal inference procedures. This involves estimating parameters of a given statistical model, and then deriving theoretical properties of the distributions of these estimates to do hypothesis testing. In contrast, machine learning is often less concerned about the "true" model that generates the data, and instead seeks out procedures that simply work well under some metric, such as predictive accuracy. This distinction is not black and white. For example, some (particularly Bayesian) machine learning algorithms begin from an assumed probability model for the data much like in econometrics, and these can in principle be used for inference. Even in these cases, though, the machine learning literature is typically less concerned with theoretical inference guarantees than is the

econometrics literature. Breiman (2001) provides a good introduction to these "two cultures" of statistical modelling.

Another area of difference is computation. Econometric procedures are rarely assessed in terms of their computational complexity, whereas such considerations are at the heart of much of machine learning. Certain core algorithms are popular precisely because they are fast to compute and can scale well. This is largely due to the massive datasets that are used in many machine learning applications. Economists can afford to work with computationally inefficient algorithms given the much smaller datasets they typically analyze, but this will evolve as datasets grow.

There are also some semantic differences that sometimes obscure what are in fact similar ideas. Both fields write models that relate some variable of interest, denoted y, to some other variables potentially related to y, denoted x. Econometricians usually call y a "dependent variable", or "outcome" and the x variables "covariates", "explanatory variables", or "independent variables". In machine learning y is often called a "label", "response", or "target", while x are "features", or "predictors". Moreover, the process of building a model to relate x and y in econometrics is called "estimation" and in machine learning "learning". This chapter will adopt the standard language of econometrics.

Rather than debate the exact definition of machine learning, it is helpful to consider instead the specific tasks that machine learning is designed to solve. A typical division is between *supervised learning and unsupervised learning,* which we now turn to discuss.

1. Supervised Learning

Supervised learning is the task of building a model to explain an outcome variable given covariates. This is exactly what many econometric models do, but the metric to judge the quality of a model in machine learning is quite distinct. Essentially, the only goal is predictive accuracy. Achieving high predictive accuracy with a fixed dataset is trivial. A linear regression model in which one uses as many covariates as there are observations to explain an outcome will perfectly explain the data. Procedures like these, however, tend to over-fit the data and make predictions based on spurious relationships. Machine learning therefore targets *out-of-sample* predictive accuracy. The goal is to build a model that accurately predicts outcomes in new data that was not used in the building of the model in the first place. Models that are good at this task are deemed successful.

To take a concrete example, consider the case of spam email. The outcome variable is binary: either an email is spam or it is not. The covariates are the words in emails. Given some fixed set of emails, predicting spam is potentially as trivial as finding a single word that is only present in spam emails and never in non-spam emails. Suppose this word is "xxx". Then, the presence of "xxx" is a perfect predictor of spam in this specific set of emails. But this model may not generalize well to new emails, for example spam emails requesting bank account details to receive the sender's inheritance. Instead, we want a model that is likely to accurately classify *new* emails as spam or not.

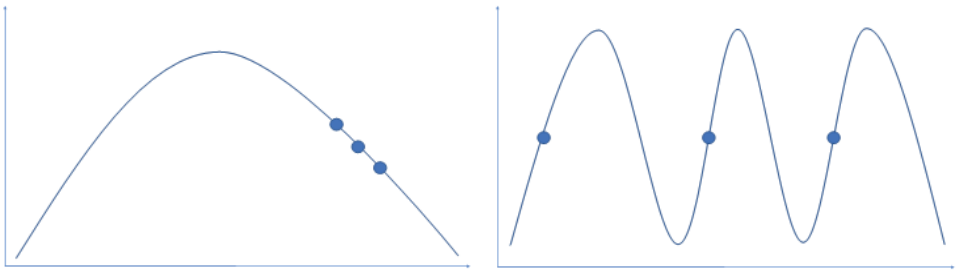
The machine learning literature has made enormous strides in building models with good predictive accuracy. Algorithms for face recognition in photos, speech recognition, and the aforementioned spam detection problem are now widely used across society, and are all applications of supervised learning.

Even if one wishes to use predictive accuracy as the benchmark to judge the success of a model, there are concerns with whether the way supervised learning algorithms are evaluated is sufficient. How can one evaluate the performance of an algorithm on out-of-sample data if such data is not available? The standard solution is to divide the data into two portions: a training sample and a test sample. The training sample is used to estimate a model. Then, for each observation in the test data, one can generate a predicted value for the outcome given the model estimated with the training sample, and then compare the prediction against the actual value in the test data. The test sample stands in for out-of-sample data since it is not used in training. However, there is often no guarantee that the *actual* out-of-sample data that an algorithm will confront in the real world corresponds to the data it confronts in the test set. Figure 1 below provides an illustration. Consider the situation on the left. Suppose the observed data are the three points on the curve, and we are trying to predict an outcome measured on the vertical axis given some covariate measured on the horizontal axis. The curve represents the real-world relationship between the covariate and the outcome. A supervised learning algorithm constructed only with the three observed points might go badly wrong *even if* it achieves high out-of-sample predictive accuracy on a test set. This is because all the data comes from a restricted part of the curve that behaves like a downward-sloping line, and a supervised algorithm will tend to estimate just this pattern. This pattern clearly does not generalize well to all possible covariate values since part of the real-world relationship involves an upward slope. Similarly, in the situation on the right the observed data again will give a misleading view on the true relationship. The problem is now that the observed data are too dispersed.[2]

[2] Many thanks to Bryan Pardo of Northwestern University who first made these points to the author.

FIGURE 1

DANGERS OF UNREPRESENTATIVE DATA FOR SUPERVISED LEARNING EVALUATION



These examples are simple and involve a single-dimensional covariate. In real applications of machine learning, one has hundreds, thousands, or even millions of different inputs, and determining whether the data on which supervised algorithms are evaluated gives a representative view of the world is extremely challenging. Economists and policymakers should bear this in mind. While mistakes in speech or image recognition can be annoying and embarrassing, they have low social costs. Mistakes in policymaking can be catastrophic.

Supervised learning algorithms are also generally constructed in environments that are quite different than those that economists face. First, they are data rich. Companies like Facebook and Google can draw on vast troves of data to train recommendation algorithms. In contrast, economists many times have very limited data to work with. For example, while predicting recessions is an important policy problem, recessions are relatively infrequent in historical time series. Second, the environments are stable in the sense that the future looks much like the past. Economies are often non-stationary, and often when predictive accuracy is most important, such as at the onset of a financial crises or the introduction of a disruptive technology. This brings into question whether off-the-shelf machine learning methods are appropriate for the kinds of prediction problems that economists are most interested in. The chapter returns to this issue in the discussion of applications below.

To better understand the differences between machine learning and traditional econometrics, it is instructive to consider a popular supervised algorithm called the LASSO (Least Absolute Shrinkage and Selection Operator), which was introduced by Tibshirani (1996) and has become increasingly popular in economics (see, for example, Belloni, Chernozhukov and Hansen, 2014). The

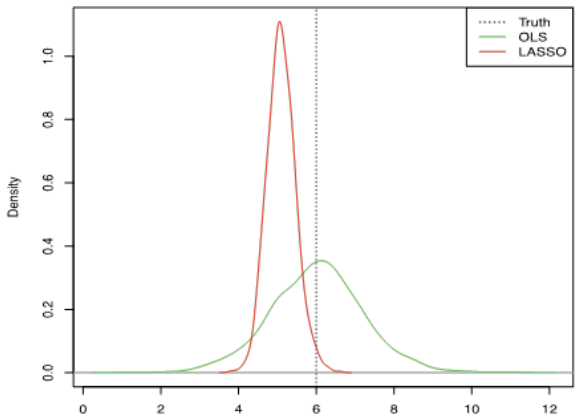
LASSO is a basic extension of the ordinary least squares (OLS) regression model that is the workhorse for applied economics. Both models relate an outcome y to covariates x by choosing coefficients for the x values that best explain y. For example, y might be income, and x might be composed of three variables: years of schooling, IQ, and eye color. We expect the first two covariates to relate to income, but the not third. The key difference between OLS and LASSO is that LASSO adds a penalty term that punishes large coefficient values. The idea behind this penalty is to assign a zero coefficient to unimportant variables and a non-zero coefficient to important variables. The hope is that the variables with non-zero coefficients have a true relationship with the outcome, and those with zero coefficients are noise variables that do not. In the previous example, this would mean that LASSO would give a positive coefficient to schooling and IQ, and a zero coefficient to eye color. Such an approach can be particularly fruitful when there are many variables relative to the number of observations. In fact, LASSO can even be estimated when there are many more variables than observations.

While the penalty term in LASSO can eliminate noise variables, this comes at a cost. The penalty term punishes large coefficient values for *all* covariates. This means that even the coefficients on the true variables are lower than they would be in the simple OLS model. In the technical language of econometrics, the coefficient values estimated from LASSO have a bias: the estimated effect of any covariate has on average a lower magnitude than whatever the true effect is. To continue with the example above, suppose that an extra year of schooling leads to an extra income of 600 EUR per year. The LASSO might estimate that the extra effect of a year of schooling is only 300 EUR per year. Why, then, would one want to use a model that intentionally introduced bias into its estimation procedure? The answer is that introducing bias reduces noise. The OLS model will estimate some coefficient value for eye color even though this has no relationship to income. On average this will be close to zero, but depending on the randomness in any specific dataset there may be some spurious correlation between eye color and income that OLS will pick up. This in turn introduces noise in predicted income. By contrast, LASSO will simply tend to drop eye color out of the model completely.

Figure 2 illustrates these properties. Suppose there is a person who has been to school for 5 years and has an IQ of 100. Moreover, suppose that an additional year of school increases income by 0.6 units, and an additional point of IQ increases income by 0.03 units. Thus, this individual's true income is $5 * 0.6 + 100 * 0.03 = 6$. Figure 2 plots the distributions of the values for predicted income produced by OLS and LASSO when there are also many noise variables in the model with no relationship to income. Here we see clearly what is known in the machine learning literature as the *bias-variance tradeoff.* OLS on

FIGURE 2

THE TRADEOFF BETWEEN BIAS AND VARIANCE



average generates the correct prediction since the distribution is centered at 6. But around this average we see large dispersion: there are predicted values as low as 0 and as high as 12. By contrast, LASSO is biased, since the distribution is centered around 5 rather than 6. But the predicted values are tightly centered around 5, without the extremes of OLS. Put another way, LASSO is wrong on average, but never too wrong; OLS is right on average, but often very wrong. One can show in this example that the average squared error–a popular metric for goodness-of-fit–is lower under LASSO than OLS.

What are implications of this example? Much of textbook econometrics restricts attention to models that are on average correct (unbiased), and then searches within such models for those with low variance. Machine learning shows us that this approach may be limited, especially when there are many variables and when the main goal is prediction, in which case biased models can perform well. At the same time, as discussed above, economists are interested in models with good inference properties: when deciding the amount to invest in public schools, it is crucial to know the true effect of an additional year of school on income (0.6 in the example above). Since supervised learning algorithms are designed for predictive accuracy, a natural question to ask is whether the two goals are in tension. In other words, can supervised learning algorithms be used for parameter inference even though they were not designed with this goal in mind? In many important cases the answer is "no", or perhaps more accurately, "not without modification". As we have seen for the LASSO, the coefficient estimates have a downward bias. Moreover, there is no guarantee that LASSO

omits all noise variables. There is some theoretical work on statistical inference with the LASSO (interested readers can consult Bühlmann and van de Geer, 2011 or Hastie, Tibshirani and Wainwright, 2015), but in practice there are few reliable guarantees that are consistent across applications.

The main message here is that supervised learning has recently made enormous strides in accurate out-of-sample prediction in stable, data-rich environments. It often does so by introducing bias to reduce variance, which is crucial in models with vast numbers of variables. However, whether and when these models can be used for the inference problems many economists care about is still an open question undergoing active research. We will discuss recent contributions in the applications section below.

2. Unsupervised Learning

While unsupervised learning has received somewhat less attention in the literature, it is important in its own right. The goal of unsupervised learning is to uncover hidden structure in data. There is no notion of a dependent variable in unsupervised learning that one tries to explain with covariates. Each observation in a dataset simply has multiple recorded variables with potentially complex interdependencies that unsupervised learning tries to reveal. There may be several motivations for unsupervised learning. One may wish to describe the most prominent sources of variation within a vast array of covariates. Alternatively, unsupervised learning can provide a low-dimensional representation of a high-dimensional object that preserves most of the relevant information. Unsupervised learning can also group observations together based on similarity. None of these motivations should be wholly unfamiliar to economists. Clustering and factor analysis, for example, are examples of unsupervised learning tasks that are already rather common in empirical economics.

Unsupervised learning can be an end in itself if data exploration is the primary goal, or else be seen as a data preparation tool used to extract features to serve as inputs into supervised learning algorithms or econometric models. In applied economics research, this makes it arguably less controversial than supervised learning. For better or worse, even in economics issues of formal inference are often downplayed when the primary aim is data processing and preparation. In this sense, off-the-shelf methods for unsupervised learning can be applied more readily if they provide a richer description of data than existing methods. The rest of the discussion argues this is indeed the case.

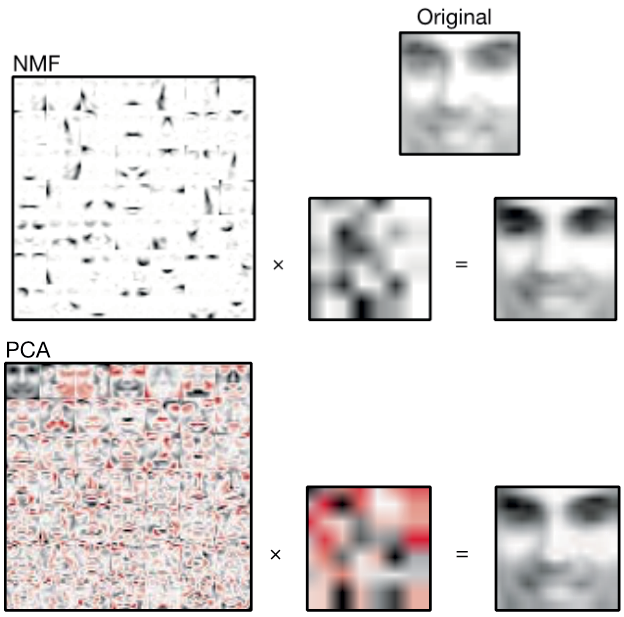
Probably the most well-known unsupervised learning algorithm in economics is principal components analysis (PCA). The idea is to find common

components across variables that explain how they move together. Observations are then represented as combinations of these common components rather than in terms of the original variables. Researchers typically use far fewer components to represent observations than there are variables, so there is dimensionality reduction. For example, such an approach is often used in macroeconomic time series to explain the co-movement of hundreds of different economic indices. The common components can be thought of as unobserved cyclical variables that drive the observed data.

PCA is also well-known in the machine learning literature, but machine learning has also developed additional algorithms that correct for some of PCA's limitations. Although economists are not generally aware of these, incorporating them into the econometric toolkit can be done at fairly low cost. One limitation of PCA is that the components it identifies can be difficult to interpret, and, in many instances, appear more like abstract objects that explain co-movement rather than objects with actual meaning. There has been work on alternative ways of constructing components in the machine learning literature that eases this problem in certain applications. An interesting example is from Lee and Seung (1999), who compare PCA with an alternative called non-negative matrix

FIGURE 3

PRINCIPAL COMPONENTS ANALYSIS VS. NON-NEGATIVE MATRIX FACTORIZATION



Source: Lee and Seung (1999).

factorization (NMF). NMF is similar to PCA, except it constrains the components to be made of only non-negative numbers. This seemingly technical distinction is in fact is substantive, because the components that NMF produces appear more like the elemental parts that each observation in the data is built from.

Figure 3 illustrates this idea for image data. The underlying dataset is a collection of photographs of human faces. The seven by seven, larger matrices on the left of the figure illustrate the 49 components that PCA and NMF uncover from the photos. Black shading indicates positive numbers, and red shading indicates negative numbers. The fascinating aspect of the example is that the NMF components appear to be elements of a face: there are eyes, mouths, noses, etc. A single photo in the data is then built by combining these elements into an individual face (the smaller matrices in the middle of the figures show the picture-specific weightings applied to the components to arrive at the observation on the right). The components of PCA are very different: the first component is essentially an average face, and the rest of the components add and subtract pixel intensity from this average face. A specific face is then represented as a weighted deviation from the average face, which is a less intuitive construction than NMF gives.

This example may seem like a curiosity, but it illustrates a deeper point that economists could potentially gain insight on latent structure from leveraging common algorithms in machine learning that have to date been almost entirely ignored. For example, one could apply NMF and related algorithms to individual product sales across consumers to learn archetypal shopping patterns and identify substitutes and complements, or to individual product prices to learn the underlying components in overall inflation.

Another limitation of PCA is that its foundations are most appropriate for data that varies continuously. One important example of data for which this is not the case is text. The most basic way of representing textual databases, also called corpora, is to count the occurrence of all unique terms in the vocabulary across all documents. The resulting data clearly has interdependencies, for example the word ‘labor’ will tend to co-occur with the word ‘wage’. But the data is fundamentally discrete, as a word cannot appear 1.5 times. Also, the vast majority of unique words in corpora do not occur in any specific document, and so the data is also populated by a large percentage of zeros. Such data calls for algorithms that model its specific features.

One of the most powerful and popular unsupervised learning models for text is Latent Dirichlet Allocation (LDA), introduced by Blei, Ng and Jordan, (2003). LDA is an example of a probabilistic topic model, which both identifies topics in corpora and then represents documents as combinations of those topics. More specifically, a topic is a probability distribution over all the unique

words in the corpus. This probabilistic aspect of LDA is important. Suppose one imagines a topic about inflation and another about unemployment. Now consider the word 'rate.' *Prima facie* it is unclear into which topic 'rate' should go, since a topic about inflation or unemployment might feature 'inflation rate' or labor 'participation rate', respectively. Allowing probabilistic assignment of words to topics conveys this semantic flexibility. LDA is also a mixed-membership model because documents are not assigned a single topic. Instead, each document is allocated shares of all topics. So, a document can be 25% about unemployment, 10% about inflation, etc.

Figure 4 shows example output of LDA estimated on a corpus of verbatim transcripts of the discussions of the Federal Open Market Committee, which decides on monetary policy in the United States. The sample period for estimation is 1987-2009. The two word clouds represent two different estimated topics. The size of the word in the cloud is approximately equal to its probability in the topic.[3] Although the algorithm is not fed any information on the underlying content of the data, the topics are clearly interpretable: there is one about economic growth, and another about recession and recovery. The time series above the topics shows variation in the share of time that individual FOMC members spend discussing the respective topics (the blue dash is the maximum share in a given meeting, the solid black line is the median share, and the dashed red line is the minimum share). Periods of recession are shaded in gray. The series also shows very natural properties. Attention to growth systematically increases when the economy expands, then collapses at the beginning of recession periods. In contrast, attention towards recession spikes during contractions. Again, it is worth emphasizing that such patterns have been wholly captured by a machine learning algorithm, with no input from the researcher.

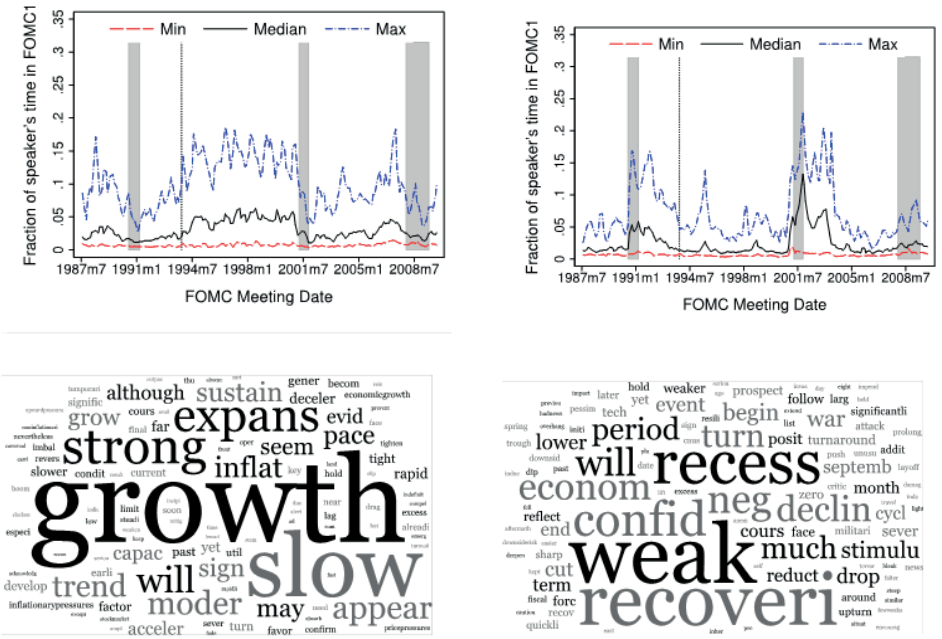
Another important point is that text is innately very high-dimensional. Even moderately-sized corpora contain thousands of unique terms. Overfitting such data is a serious problem, but the statistical structure of LDA guards against this. It is what is known as a Bayesian model, which means it places some initial likelihood on all possible combinations of words in topics. The observed data then changes these likelihoods but does not fully determine them. The transcript dataset above has roughly 10,000 unique terms, and yet LDA handles the dimensionality with ease.

These two examples show the power of unsupervised learning to reveal interesting patterns in data. Moreover, they also show how machine learning can convert what at first sight are unstructured, messy data–*i.e.,* image files and raw text data–into a tractable, quantitative forms that are suitable for

[3] Some of the terms are not English words because the data has been stemmed prior to estimation, a process whereby words are brought into their linguistic roots.

FIGURE 4

EXAMPLE OUTPUT OF LATENT DIRICHLET ALLOCATION



Source: Hansen, McMahon and Prat (2018).

traditional statistical analysis. This opens the possibility of not just having new techniques to use with existing data, but having access to new data itself. This point is discussed further in the applications section below.

One possible criticism of unsupervised learning algorithms, however, is that they have too little structure. Figure 4 shows that there are likely to be time-varying probabilities of topic coverage depending on the business cycle, but this is not built into LDA. One possible contribution of economists to the development of unsupervised learning algorithms is to introduce dependencies of interest into them to more directly link their outputs to quantities of interest. Such efforts will likely require collaboration across disciplines.

III. APPLICATIONS

Having set the foundations of basic concepts in machine learning, the rest of the chapter expands on potential applications in economics and policy. We

begin with one of the most pragmatic applications: to quantify novel data in tractable forms. Next, we consider the role of machine learning in converting digital data into specific economic measures, followed by a discussion of machine learning in forecasting models. Finally, we reflect on possible applications in causal inference.

1. Quantification of Unstructured Data

Many firms and regulators are awash with unstructured data, and specifically text data. One leading example is the legal industry, in which much of the work of junior lawyers is taken up by trawling through documents to find relevant content from contracts, title deeds, prior judicial decisions, etc. Regulators too face a similar task when they initiate cases. For example, dawn raids on potential violators of competition law typically yield troves of documents, and sifting out the relevant material from the mass of irrelevant material is an important challenge. Automating the task of finding relevant information therefore has the potential to generate large efficiency gains in these contexts, and indeed this process is already well underway in the legal industry (Croft, 2017).

One of the most common ways of determining document relevance in economics is keyword searches. In this approach, a word or list of words is defined in advance, and then documents are flagged as containing these terms or not, or alternatively ordered according to the frequency with which terms appear. While simple and relatively easy to implement, keyword searches have limitations. Most basically, they require the definition in advance of the important words, which may require subjective judgments. For example, to measure economic activity, we might construct a word list which includes 'growth'. But clearly other words are also used to discuss activity, and choosing these involves numerous subjective judgments. More subtly, 'growth' is also used in other contexts, such as in describing wage growth as a factor in inflationary pressures, and accounting for context with keyword searches is practically very difficult. In other cases, the academic or policymaker may simply have no idea how words relate to the content of interest. In litigation involving traders' manipulation of market prices like the recent LIBOR rate-fixing scandal, much of the evidence comes from chat rooms in which traders make heavy use of jargon, slang, and code that make simple keyword searches difficult to implement.

Unsupervised machine learning helps overcome some of these problems. Especially in environments with uncertainty about what content documents contain, and how words are used in different kinds of contexts, machine learning provides a powerful, data-driven approach for corpus exploration and information retrieval. The quantification of unstructured data might be an end in itself by, for example, allowing a regulator to quickly sift through documents

and sort them into categories. Or it might be the first stage in extracting features from text data that then serve as inputs into further empirical studies.

To illustrate these points more concretely, consider the example data point in the FOMC transcript corpus discussed in the previous section represented in Figure 5. This is an utterance of Janet Yellen in March 2006 when she was President of the Federal Reserve Bank of San Francisco. This statement uses highly technical language, and determining its content manually would require a reader to have a high level of education in economics.

As an alternative to manual processing, one can use Latent Dirichlet Allocation (LDA), an unsupervised learning algorithm described above, to

FIGURE 5

EXAMPLE DATA POINT IN FOMC TRANSCRIPTS

> We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

Source: Janet Yellen, March (2006).

determine its content. The estimated LDA model associates this statement most with the topic on the left in Figure 6 below. This topic in turn places highest probability on 'inflation' (and other words that begin with the stem 'inflat'). The fascinating aspect of this illustration is that the example data point contains no occurrence of the word `inflation', and a keyword search for it would not flag this statement as relevant. Instead, Janet Yellen uses many words related to inflation (CPI is the consumer price index, PCE is personal consumption expenditure), and LDA learns from other documents in the corpus that the words Yellen uses are most often used in situation in which the word `inflation' is also used. This allows it to associate her statement with the inflation topic.

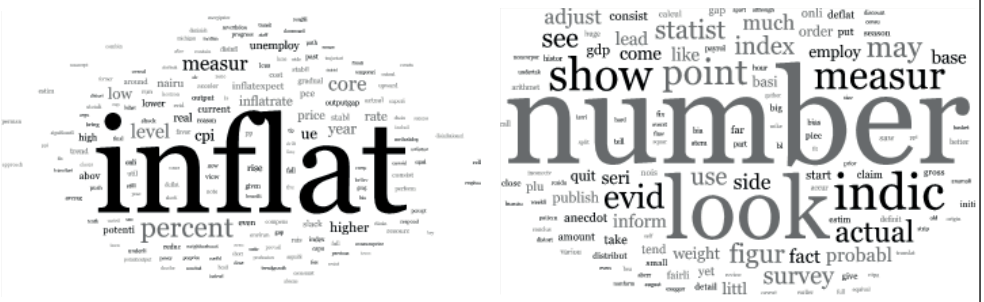
Another point of interest is that LDA is able to place individual words within documents into their appropriate context. Consider the word 'measures' that Yellen uses in the example statement. While this word appears prominently in the inflation topic, it is also present with high probability in another topic about numerical indicators displayed on the right in Figure 6. LDA can resolve this ambiguity by looking at the other words that Yellen speaks. While the

classification of 'measures' without any context is unclear, the fact that Yellen uses many words unambiguously associated with the inflation topic causes it to assign 'measures' to the inflation topic as well.

These features of LDA help explain its widespread popularity. One application familiar to many readers might be the indexing system in JSTOR, the popular repository for academic papers. LDA is used to alert readers to new articles of potential interest in the repository given the estimated content in previously viewed articles. The adoption of such systems for firms and regulators that also handle large textual corpora would potentially create large gains in automated information retrieval.

FIGURE 6

TOPIC MOST ASSOCIATED WITH EXAMPLE (LEFT) AND ALTERNATIVE TOPIC (RIGHT)



While the focus of the discussion so far has centered around text, similar points can be made for other kinds of unstructured data too. Policy institutes and marketing firms regularly collect survey data to measure attitudes, behaviors, and characteristics. This data is often analyzed in *ad hoc* ways, for example by computing some average response across a range of questions to obtain a single number. Again, unsupervised learning provides a way of modeling the full dependency structure in the data and extracting novel insights into the underlying ways in which respondents differ from each other. One example in the economics literature is Bandiera *et al.* (2017), who analyze detailed time-use surveys of over 1,000 CEOs across a range of countries. Using LDA, they find a novel behavioral distinction between CEO 'leaders' who spend time coordinating high-level functions in companies and `managers' who spend time on more operational matters. Similar approaches have been used to measure health status (Erosheva *et al.,* 2007) and political ideology (Gross and Manrique-Vallier, 2014) from surveys.

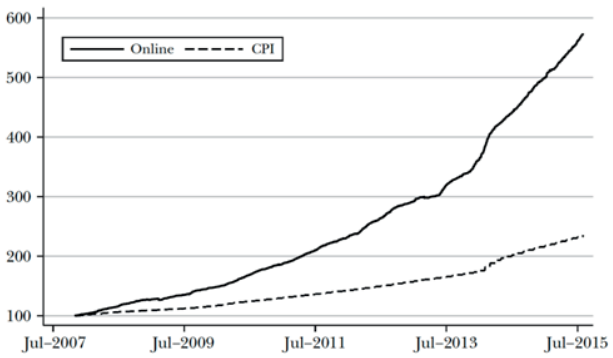
Another intriguing potential application of unsupervised learning is to network data, where the challenge is to identify groups of related nodes based on linkage patterns. There is a large literature on this so-called community detection problem outside economics, but hardly any economic applications. One exception is Nimczik (2017), who estimates the geographic extent of labor markets using data on worker flows in Austria using unsupervised learning.

2. New Data and New Measurement

The first application discussed above was simply to use machine learning to make sense of messy, difficult-to-interpret data while imposing minimal structure on the process of information retrieval. However, there is growing interest in not just describing such data, but also in using it to construct new measures of relevant economic variables. There are various ways in which traditional economic indicators are limited. They are often available at relatively infrequent intervals, as is the case with quarterly GDP measures. Furthermore, they are often constructed for aggregated geographical units like nation states with very little spatial granularity. Finally, in many regions of the world official economic statistics are either unavailable entirely, or else manipulated by governments to the extent that they contain very little information. For these reasons, there is demand for new sources of information. Recently there has been growing interest in digital data as a means of filling these gaps. Examples include:

FIGURE 7

ARGENTINIAN INFLATION AS MEASURED BY ONLINE PRICES AND OFFICIAL CPI MEASURE



Source: Cavallo and Rigobon (2016).

- In Argentina, the government actively manipulated official price statistics beginning in 2007. The Billion Prices Project at the Massachusetts Institute of Technology began as a means of providing an alternative, more accurate inflation index using prices posted by online retailers in Argentina and has since expanded to many other countries. While the universe of retailers for which one can obtain online prices is smaller than that surveyed by official government agencies, these prices are updated daily, have a low cost of extraction, and are free from government interference. Figure 7 below shows inflation measures using online prices and official statistics, and demonstrates the ability of digital data to capture the actual underlying dynamics in an economy when official data is unavailable or unreliable.

- Baker, Bloom, and Davis (2016) construct a popular and influential Economic Policy Uncertainty (EPU) Index (http://www.policyuncertainty.com/). While the impact of uncertainty on economic activity is acknowledged as important, historically there have been very few adequate measures of uncertainty. Financial-market based measures like VIX are based on option prices derived from US equity markets, which do not capture the full uncertainty that economic agents face. The EPU index instead measures uncertainty specifically about policymaking. It is constructed in large part based on the fraction of articles in a wide selection of newspapers that contain terms like 'uncertain', 'economic', 'congress', and 'regulation'.

- Glaeser, Kim, and Luca (2017) construct a local activity index using the number of restaurants and businesses reviewed on the website Yelp. This index has predictive power for the much more aggregated and lagged data from the US Census Bureau on county business patterns, especially in more densely populated areas.

- SpaceKnow is a commercial company that produces numerous indices of economic activity using satellite image data. One such index is the China Satellite Manufacturing Index, which is based on 2.2 billion individual snapshots of more than 6,000 industrial sites in China (Wigglesworth 2018).

While activity indices are some of the most natural objects of interest that new data can provide, there are also less obvious but equally powerful possibilities. A good example comes from the work of Hoberg and Phillips (2010 and 2016) and has direct relevance to competition policy. The issue is how to measure the industry classification of firms. The often-used SIC or NAICS classification systems have several limitations. Firms typically do not

receive different classifications over time even when their markets evolve. The classification systems also do not track the development of entirely new products particularly well. More generally, they provide a very coarse distinction of the ways in which firms differ from each other.

Hoberg and Phillips propose the use of text data to construct industry classifications that overcome some of these challenges. The idea is to use companies' product descriptions contained in their annual 10-K filings to the US Securities and Exchange Commission. For each pair of firms that make a filing in each year, one can compute a measure of linguistic similarity between descriptions and use it as a proxy for proximity in product space. Moreover, from these similarity measures, one can group firms into clusters to define industry categories. The resulting categorization provides a dynamic, continuous measure of firms' location in product space relative to all other firms in the data. Hoberg and Phillips show that their text-based categorization provides several new insights into why firms merge and how new products develop.

At this stage, it is useful to make the distinction between `big data' coming from digital sources on the one hand and machine learning on the other. While raw digital data no doubt contains information relevant for economic variables of interest, the exact mapping between the two is difficult to know. One possibility is to apply unsupervised learning algorithms to describe the data along the lines discussed in the first application, and then use the extracted features to build an index of interest. The problem is that these features will not have been chosen to have maximum predictive power for the economic variable, which implies a loss of information and thus usefulness.

Instead, the task of building new indices from vast data is in many ways a classic supervised learning problem, since the primary goal is to make the best possible prediction of the object of interest. Jean *et al.* (2016) is an example of research that combines vast digital data (satellite images) and state-of-the-art supervised machine learning algorithms to provide a new economic measurement (spatially granular poverty levels in several African countries). As the use of machine learning in economics becomes more widespread, many of the indices built from digital data will likely also be the output of targeted supervised algorithms.

3. Forecasting

As discussed above, supervised machine learning is at its heart the study of methods for achieving good out-of-sample prediction using high-dimensional or unstructured data. One area of high interest for policymakers is forecasting,

or predicting the future based on past data. In fact, the idea that rich economic time-series data can be used to obtain better forecasts of the future predates the growth of interest in machine learning. Stock and Watson (1999) and Bernanke, Boivin and Eliasz (2005) are seminal contribution in the literature that show that augmenting standard macroeconomic forecasting models with many time series can improve future forecasts. These papers use methods like penalized regression and dimensionality reduction that are part of the standard machine learning toolkit.

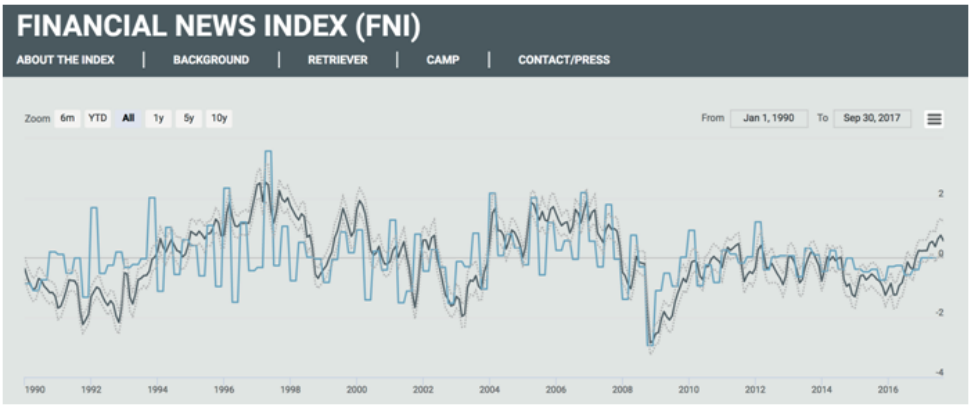
Before economists apply more modern supervised learning algorithms for forecasting, it is worth emphasizing again that the problem of economic forecasting differs in fundamental ways from the environments in which many machine learning algorithms are built and evaluated. First, a common assumption in machine learning is that the out-of-sample data has the same distribution as the training data. In a time-series context, this boils down to an assumption that the future looks like the past. While this may sometimes be true, in other cases it might not be if there are fundamental structural changes. For example, if there is shift in the productive capacity of the economy, then the historical relationship between unemployment and wage growth will change. While there is a well-established literature in econometrics on the detection of structural breaks, the machine learning literature in this area is much less developed. Second, in economics the data is often big on some dimensions but small on others. While there are hundreds of available time series for forecasting, many are observed only at a quarterly or even less frequent basis. Third, the so-called `signal-to-noise' ratio in economic and financial data can be quite low, which means that fundamental relationships among variables can be hard to detect because there is a lot of randomness that affects all variables in the model. The overall challenge, then, is to find ways of employing supervised learning methods in situations for which they were not originally designed.

One possibility is the use of so-called *generative* models. These models construct a full statistical model for input and output data, in contrast to some recent tendencies in machine learning like deep learning that take a more agnostic stance on the model that generates the data. The main reason for the success of deep learning models is their remarkable predictive power in the presence of vast data. In smaller samples like the ones economists face, though, generative models have been proven to have better predictive power (Ng and Jordan, 2002). Davig and Smalter Hall (2017) make use of this insight, and show that a generative model better predicts US recessions than standard regressions models and the Survey of Professional Forecasters. Another advantage of generative models is that they are closer to the kinds of structural models that economists are already used to constructing and estimating.

Another approach to forecasting with large data is to first use unsupervised learning to extract features, and then use those features as inputs into an otherwise standard economic forecasting model. One example is Thorsrud (2016), who applies latent Dirichlet allocation to Norwegian media articles, and uses the extracted topics to predict evolution in the business cycle. Figure 8 below plots the derived index against actual Norwegian GDP. Clearly the two series co-move substantially, which illustrates the value of features extracted from unsupervised learning for forecasting.

FIGURE 8

FINANCIAL NEWS INDEX (BLACK) AND NORWEGIAN GDP GROWTH (BLUE)



Source: From https: //www.retriever-info.com/fni

Another example from outside macroeconomics is the prediction of conflict, which is important both for risk management of private sector companies and governments. Mueller and Rauh (2017) show that media data can help forecast the outbreak of political violence. They also use LDA to extract topics from text, and then show that variation in topic usage in newspapers' coverage of countries predicts conflict in those countries.

A general comment that applies to the approach of using extracted features as inputs into forecasting models is that they implicitly treat them as fixed data rather than estimated objects. While this has led to important advances in research, in the future one would expect the development of algorithms that jointly model high-dimensional data and whatever variable is being predicted. This is likely to lead to even better predictions, and also more rigorous statistical inference. Again, generative models can provide the backbone for such approaches.

4. Causal Inference

The applications discussed so far all represent important steps in empirical work in economics, but the profession is currently dominated by interest in causal inference, and more precisely in determining the effect of policy interventions. The usefulness of predictive models for this goal is not immediately obvious. Athey (2017) presents a nice illustration of this point. Suppose a hotel chain is interested in determining the effect on sales of rooms following an increase in the price of rooms. If one simply takes observed price and sales data, there is a positive relationship because as occupancy rates increase hotels raise the price of remaining rooms: during peak holiday periods rooms are scarce and prices are high, while during low season the reverse is true. Therefore, a purely predictive model would indicate higher sales following an unexpected increase in price. Of course, common sense dictates that exactly the reverse would occur, *i.e.,* a hotel would sell fewer rooms if it unexpectedly raised prices. The problem here is that a pure predictive model based on observed data fails to account for the unobserved underlying demand for hotel rooms. High occupancy rates are associated with high prices because high demand drives both. Methods for solving problems such as these have been the subject of a great deal of modern econometrics.

What, then, can machine learning offer for economists interested in estimating causal relationships? One important realization is that even causal inference procedures involve what are essentially pure prediction steps. One classic approach for causal inference is the use of so-called 'instrumental' variables. These are variables that are correlated with a treatment but not with the outcome of interest.[4] Replacing the treatment with the instruments allows one to isolate the causal impact of the treatment on the outcome. Instrumental variable estimation typically proceeds in two steps: first, one predicts the value of the treatment given the instruments; second, one uses the predicted value of the treatment as an independent variable in a regression on the outcome. The first step in this procedure can be viewed as a natural machine learning task as it involves making an optimal prediction of the treatment given the instruments. Machine learning methods for instrumental variables are particularly relevant when there are many potential instruments, or when one wants to estimate a flexible relationship between instruments and treatments. Several recent papers combine supervised machine learning methods with instrumental variables (Belloni *et al.,* 2012; Hartford *et al.,* 2017).

Another application of machine learning to causal inference is the problem of high-dimensional controls. Many potential *observable* variables can

[4] In the following discussion, a treatment will mean a variable that a researcher or policymaker intervenes to change, and an outcome will mean whatever target variable he or she is attempting to influence.

also affect the outcome of interest beyond just the treatment of interest. For example, the impact of worker training on productivity might depend on worker characteristics, firm characteristics, and the characteristics of the technology that the worker operates. Which control variables beyond the treatment to include in regression models is often unclear, especially in the absence of a relevant theory. A common approach is to run many different models, each of which includes different controls, and to examine how sensitive the relationship between a treatment and outcome is to the inclusion of a particular set of controls. One naïve machine learning approach would be to include all controls along with the treatment in a penalized regression model in order for the data to reveal which controls are relevant. In fact, this approach yields unreliable estimates of the treatment effect, but adjustments of off-the-shelf algorithms can help correct the problem (Belloni, Chernozhukov and Hansen, 2014).

Another approach to causal inference in economics is so-called structural modelling in which one takes a theoretical economic model, and then uses data to estimate the parameters of the theory. As models grow in complexity, the number of parameters can grow rapidly. For example, a consumer demand model could in theory involve cross-price elasticities between every possible pair of goods in a supermarket. Machine learning can also offer techniques for parameter estimation in large-scale structural models fit on large-scale data. Generative models with a Bayesian formulation again provide a natural framework for structural estimation in economics. While these have arguably lost favor in recent years in the machine learning community due to the rise of deep learning, their future in economics is promising. A recent example is Athey *et al.* (2018), but it is safe to say that this application of machine learning is probably the least developed of all those discussed.

As with the forecasting application, the broad point again arises that the context in which machine learning algorithms are often built is not necessarily directly applicable to empirical applications. This is not to say that machine learning has no relevance to causal inference, but in this area especially careful thinking is required to assess where machine learning techniques can add value.

IV. CONCLUSION

This chapter has reviewed basic concepts in machine learning and provided numerous examples of how machine learning might be useful to academic economists and policymakers. Some applications simply require off-the-shelf methods, while others require the development of new techniques to address the challenges specific to economics. While some of these techniques are already under development, there is much still to be done.

While this chapter has focused on the value policymaking authorities can derive from applying machine learning techniques to data, there are also new regulatory issues that are byproducts of the increased use of machine learning. One example is firms' use of pricing algorithms. When firms tailor prices to individual customers' traits and behavior, price discrimination almost necessarily increases. Whether this reduces consumer surplus is less clear. On the one hand, increasing prices while keeping quantity constant reduces surplus, but on the other pricing algorithms may allow firms to increase the quantity or variety of goods produced. A second issue is whether the use of pricing algorithms can increase tacit collusion by providing new opportunities for firms to link their prices to the prices their competitors post. This issue is the subject of recent academic (Salcedo, 2015) and policy (OECD, 2017) interest. While there is a growing awareness of these issues, determining the appropriate responses from competition authorities is still an open question, although there is a broad understanding that "the rise of pricing algorithms and AI software will require changes in our enforcement practices" (McSweeny, 2017). Of course, addressing these questions requires at least a basic understanding of the nature of machine learning algorithms, which is another important motivation for this chapter.

Another important regulatory issue is transparency. Firms are increasingly using machine learning to automate decisions that affect consumers in important ways, but in some cases this can increase opacity relative to human decision making. One example is the decision to grant credit. Financial institutions deploy machine learning algorithms to decide which kinds of consumer receive which types of loans, but consumers do not necessarily understand the key characteristics for predicting repayment risk. Regulators in this and other situations have a role to play in ensuring transparency and fairness.

Finally, much of the digital data valuable for machine learning applications is held by private sector companies whose main interest in exploiting it is commercial. To the extent that such data also has public value for research and policymaking, regulators will also be called upon to facilitate the transfer of data from the firms that directly collect it to a wider range of interested parties.

BIBLIOGRAPHY

Athey, S. (2017), "Beyond prediction: Using big data for policy problems," *Science,* 355: 483-485.

Athey, S.; Blei, D.; Donnelly, R.; Ruiz, F., and T. Schmidt (2018), "Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data," *American Economic Review Papers and Proceedings,* forthcoming.

Baker, S. R.; Bloom, N., and S. J. Davis (2016), "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics,* 131(4): 1593-1636.

Bandiera, O.; Hansen, S.; Prat, A., and R. Sadun (2017), CEO Behavior and Firm Performance, *NBER Working Paper,* 23248.

Belloni, A.; Chen, D.; Chernozhukov, V., and C. Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica,* 80: 2369-2429.

— (2014), "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspective,* 28(2): 29-50.

Bernanke, B. S.; Boivin, J., and P. Eliasz (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics,* 120(1): 387-422.

Blei, D.; Ng, A., and M. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* 3: 993-1022.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science,* 16(3): 199-231.

Bühlmann, P., and S. van de Geer (2011), *Statistics for High-Dimensional Data: Methods, Theory, and Applications,* Springer Series in Statistics, Springer.

Cavallo, A., and R. Rigobon (2016), "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives,* 30(2): 151-178.

Croft, J. (2017, May 4), "Artificial intelligence closes in on the work of junior lawyers," *Financial Times,* retrieved from www.ft.com

Davig, T., and A. Smalter Hall (2017), Recession Forecasting Using Bayesian Classification, *The Federal Reserve Bank of Kansas City Research Working Paper,* 16-06.

Einav, L., and J. Levin (2014), "Economics in the age of big data," *Science,* 346 (6210).

Erosheva, E. A.; Fienberg, S. E., and C. Joutard (2007), "Describing Disability through Individual-Level Mixture Models for Multivariate Binary Data," *The Annals of Applied Statistics,* 1(2): 502-537.

Glaeser, E. L.; Kim, H., and M. Luca (2017), Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity, *Harvard Business School Working Paper,* 18-022.

Gross, J. H., and D. Manrique-Vallier (2014), "A Mixed Membership Approach to the Assessment of Political Ideology from Survey Responses," in Airoldi, E. M.; D. Blei; E. A. Erosheva, and S. E. Fienberg, editors, *Handbook of Mixed Membership Models and Its Applications,* CRC Press.

Hansen, S.; McMahon, M., and A. Prat (2018), "Transparency and Deliberation on the FOMC: A Computational Linguistics Approach," *Quarterly Journal of Economics,* forthcoming.

Hartford, J.; Lewis, G.; Leyton-Brown, K., and M. Taddy (2017), Deep IV: A Flexible Approach for Counterfactual Prediction, *Proceedings of the 34th International Conference on Machine Learning.*

Hastie, T.; Tibshirani, R., and M. Wainwright (2015), *Statistical Learning with Sparsity:The Lasso and Generalizations,* Number 143 in Monographs on Statistics and Applied Probability. CRC Press.

Hoberg, G., and G. Phillips (2010), "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis," *Review of Financial Studies,* 23(10): 3773-3811.

— (2016), "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy,* 124(5): 1423-1465.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B., and S. Ermon (2016), "Combining satellite imagery and machine learning to predict poverty," *Science,* 353(6301): 790-794.

Lee, D. D., and H. S. Seung (1999), "Learning the parts of objects by non-negative matrix factorization," *Nature,* 401(21 October 1999): 788-91.

Marr, B. (2015, September), "Big Data: 20 Mind-Boggling Facts Everyone Must Read," *Forbes,* retrieved from https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#5741b5117b1e

McSweeny (2017), *Algorithms and Coordinated Effects. Remarks of Commissioner Terrell McSweeny,* University of Oxford Center for Competition Law and Policy, 22 May 2017.

Mueller, H., and C. Rauh (2017), "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text," *American Political Science Review,* forthcoming.

Mullainathan, Sendhil, and Jann Spiess (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives,* 31(2): 87-106.

Ng, A. Y., and M. I. Jordan (2002), On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes, *Neural Information Processing Systems.*

Nimczik, J. S. (2017), *Job Mobility Networks and Endogenous Labor Markets,* unpublished manuscript, Humboldt University Berlin.

OECD (2017), *Algorithms and Collusion: Competition Policy in the Digital Age,* www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm

Salcedo, B. (2015), *Pricing Algorithms and Tacit Collusion,* unpublished manuscript, Pennsylvania State University.

Stock, J. H., and M. W. Watson (1999), "Forecasting inflation," *Journal of Monetary Economics,* 44: 293-335.

Thorsrud, L. A. (2016), Words are the new numbers: A newsy coincident index of business cycles, *Working Papers,* 4/2016, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society Series B,* 58(1): 267-288.

Varian, H. R. (2014), "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives,* 28(2): 3-28.

Wigglesworth, R. (2018, January 31), "Can big data revolutionise policymaking by governments?," *Financial Times,* retrieved from www.ft.com

BITCOIN: A REVOLUTION?[1]

Guillaume HAERINGER
Hanna HALABURDA[2]

Abstract

Today more and more people talk about Bitcoin, cryptocurrencies, blockchain, or smart contracts, and many predict that these technologies will revolutionize our lives. But the apparent complexity of Bitcoin and its related technology makes it hard to participate to the debate. The purpose of this chapter is to offer a non-technical description of this new phenomenon, giving answers to many common questions (*e.g.,* the electricity consumption of Bitcoin, or the necessity of "wasteful" mining) and debunking some of the myths surrounding Bitcoin and the blockchain technology (*e.g.,* that it is 100% tamper proof).

Key words: Bitcoin, blockchain technology, wasteful mining, electricity consumption.

JEL classification: D80, G20, O30, O33.

[1] We thank Debbie Haeringer, Lukasz Pomorski, Meredith Stevens and Larry White for helpful comments and suggestions.

[2] Views presented here are views of the author and do not necessarily represent a position of Bank of Canada.

I. INTRODUCTION

In 2008, a white paper describing a cash system (*i.e.,* a "currency") that could be fully decentralized was published on the internet. Satoshi Nakamoto is the pseudonym of the author, or group of authors, who invented the system and wrote the proposal. He (or she or they) coined (pun intended) this new system *Bitcoin*.[3] Ten years later, Bitcoin became a household name, periodically making the headlines of newspapers, blogs and other media. Many would say that Bitcoin is a revolution. At least, this new "currency" is puzzling. The volatility of its exchange rate is off the charts, challenging financial analysts and scholars alike. There is a concern that the traditional tools to analyze assets are not adequate here (see, *e.g.,* Urquhart, 2016). Many central banks and regulators, wary of the consequences of such volatile and uncontrolled currency, are taking active interest in the developments surrounding it. What makes Bitcoin "a revolution"?

At first glance, Bitcoin seems revolutionary because it offers us a money system without a trusted third party. But in a way, it is going back to basics. All early money systems ran without a trusted third party. So, what's new about Bitcoin?

We tend to think of money as means of exchange governed by authorities and trusted third parties such as governments, banks, central banks, credit unions etc. But these are quite recent phenomena in the history of money (see for instance Ferguson, 2009 or Halaburda and Sarvary, 2016). In essence, money (or a cash system) is just something that a group of people use to facilitate exchange of goods and services. If we think back to the earliest examples of money, these were sea shells, animal teeth, and eventually pieces of precious metals. Such money did not need any authorities or trusted third parties.[4]

With time, coins issued by kings and emperors would use a stamp to certify the amount of metal in the coin. Such certification facilitated transactions, as it saved the time and hassle of weighing the metal with each transaction. It worked, however, only if the certification was trusted.

Metal is heavy to carry around, especially if we think beyond gold and silver. In 17th century Sweden, copper was the currency metal. Exchanging meaningful value meant moving copper plates weighing more than a few kilograms.

[3] The paper is available here: https://bitco.in/pdf/bitcoin.pdf. As of today the true identity of Satoshi Nakamoto remains unknown.

[4] In modern times, cigarettes used in PoW camps and in prisons are similarly examples of money ran without any trusted third party. See Radford (1945).

Trusted third parties like banks, and later central banks further facilitated trade by providing paper money which was more convenient to carry around and use. The value attached to paper money was initially related to the promise of the banks to redeem the notes for a specified amount of metal. But in the end, all what matters is whether a particular token is expected to be accepted in the next transaction. And so, paper money was also successful as means of exchange even after banks moved away from the gold standard and stopped relating notes' value to the amount of metal.

Paper money is more convenient as a medium of exchange than sea shells or pieces of gold. While sea shells and gold are naturally difficult to obtain (and impossible to create for the latter, although generations of alchemists have tirelessly tried), we need to trust that an authority will make paper money adequately scarce by not printing too much or too little, and by policing counterfeiting.

In this technological age, digital money comes in the form of 0s and 1s and more physically as cards and chips. It is even more convenient to use than paper money. This convenience is reflected in the decline of cash used for purchases, while credit and debit card usage is increasing. At the same time, the service of ensuring scarcity provided by trusted third parties is even more important for digital forms of money. This is because making perfect copies of digital money (*i.e.,* counterfeiting) is very cheap. Such copying would allow for spending the same digital coin more than once. To prevent such *double spending,* all digital payment systems before Bitcoin relied on a trusted third party (*e.g.,* a bank) to keep track of all the money spent and to make sure nobody can spend the same money multiple times.

The innovation of the Bitcoin system is that, for the first time, it offered a *digital money* without a trusted third party. In the next sections we describe how the Bitcoin system is able to achieve it. Before we go there, however, it is worth mentioning that constructing a fully decentralized (*i.e.,* not needing a trusted third party) system for digital currency has been a long-standing challenge in the cryptography community, dating back at least to 1980's. Previous attempts considered it a purely cryptography question and focused on cryptography solutions. But Bitcoin was the first one to succeed, and it did so by combining cryptography tools with incentive systems to prevent double-spending.

Sections II and III explain in more detail how Bitcoin works, but without entering into technicalities. Sections IV and V discuss current and potential uses of Bitcoin and technologies inspired by it.

II. THE BITCOIN PROTOCOL

Bitcoin is a digital cash system that involves a currency called *bitcoin* (with a lowercase *b*, symbol ฿) and two types of actors: the users and the miners.[5] Bitcoin is a pure digital cash system, which means that there are no physical versions of bitcoins in the form of coins or paper notes.[6]

A user is any person or entity holding or receiving bitcoins, and a miner is any person or entity that records and validates transactions. From a user's perspective the Bitcoin system may not seem too different from the banking system with bank accounts (although with very limited services: we can only make deposits and transfers). The details of how the transfers are validated and settled are different, though.

1. Wallets and the Blockchain

To become a user in the Bitcoin system, one needs to set up an account, which is a bit like having a debit card with a PIN. One can easily create a Bitcoin account (or many of them) on a website like bitaddress.org or blockchain.info. When a user creates a Bitcoin account she obtains a string of characters and a number. The string is the *Bitcoin address,* which is the equivalent of a debit card number or the bank account number, and looks like this:[7]

`12c6DSiU4Rq3P4ZxziKxzrL5LmMBrzjrJX.`

To send bitcoins to someone we need a Bitcoin address of that person, our Bitcoin address and our *Bitcoin private key.* The latter is a little bit like the PIN of your debit card. For Bitcoin that PIN is not made of 4 digits, it is a much longer number. It is 77 digits long! To make it "easier" this number is usually represented with a string of characters, similar to a Bitcoin address. Here is an example of what a Bitcoin private key looks like:[8]

`873D79C6D87DC0FB6A5778633389.`

[5] We use the term "Bitcoin" with an uppercase *B* to refer to the cash sytem. Like most currencies Bitcoin can take decimal values. The smallest unit is called the *satoshi*, and it represents one hundred millionth of a bitcoin, *i.e.,* one satoshi is equal to 0.00000001 bitcoin.

[6] The "Bitcoin coins" one can buy on websites like eBay or Amazon are not bitcoins, there are mere pieces of metal on which the word "Bitcoin" has been engraved.

[7] This Bitcoin address is one the very first Bitcoin addresses that has been created and is believed to belong to Nakamoto himself/herself.

[8] Bitcoin supports different ways to encode (*i.e.,* rewrite) the private key into a string of characters.

Together, the Bitcoin address and the Bitcoin private key form the *Bitcoin wallet.* Note that under the Bitcoin terminology we do not use the word "account" but rather "address" (or "wallet"). So now we are all set: Users have account numbers (the Bitcoin address) and a PIN (the Bitcoin private key) that allow them to spend or receive bitcoins. Note that unlike debit cards, one can have a practically *unlimited* number of Bitcoin wallets. One can create a new Bitcoin wallet for every transaction. But, if there are no banks, how are the bitcoins stored? The answer is what has now become a buzzword: the *blockchain*.

To start with, the Bitcoin blockchain is a computer file.[9] More precisely it is a ledger that records all the transactions that have ever been made with Bitcoin. Transactions in the blockchain are grouped in blocks (batches), and the sequence of blocks constitutes the blockchain. This is where its name comes from: a *chain* of *blocks* of transactions. Whenever a user sends bitcoins to another address, that transfer is stored in the blockchain.

Unlike account numbers at a bank, the blockchain is public: anyone can look into it. What the blockchain does not contain, however, are the names of the owners of the Bitcoin addresses stored in it.[10]

Remark 1. *Many people think that because the blockchain only contains Bitcoin addresses it is an anonymous form of payment. Not quite so. Computer scientists have shown that a careful analysis of the blockchain, and cross-referencing the information contained in it with other sources, may provide an opportunity to identify some users* (see Androulaki *et al.,* 2013).

If we want to know the balance of an address we have to parse the whole blockchain looking for that address, and the balance is simply the sum of all incoming transactions minus the sum of all outgoing transactions (we don't need to know the private key associated with that address). There are a number of websites doing that for us, so we do not need to download the whole blockchain and search it. The same websites also offer tools to easily send bitcoins from one address to another address.

There are tens of thousands of computers across the planet with a copy of the blockchain and that are maintained by people called miners (we describe what they do in the next section).[11] The multiplicity of copies of the blockchain

[9] A large one and growing: at the end of 2017 the size of the Bitcoin blockchain was nearly 150 gigabytes.

[10] This is why we said that the Bitcoin address shown above is believed to belong to Satoshi Nakamoto.

[11] The estimates vary from 10,000 (bitnodes.earn.com) to 30,000. See also https://en.bitcoin.it/wiki/Clearing_Up_Misconceptions_About_Full_Nodes

brings some trust to the system. Since each copy of the blockchain contains the bitcoins associated with any wallet there is no risk that they could be lost due to a computer failure. The multiplicity of copies, however, does not protect the blockchain from manipulation and fraud. We will see in Section III how other Bitcoin properties play a role in preventing that.

Remark 2. *Since Bitcoin is a system that works without third parties, holdings of bitcoins are not secure. All that thief needs is your wallet (the Bitcoin address and the Bitcoin private key). This differs from credit cards or bank accounts which usually have some insurance against theft. More importantly, if a user loses her Bitcoin private key there is no known way to recover it. The bitcoins in the corresponding wallet are gone forever. It is believed that approximately 4 million bitcoins have been lost since 2008.*[12] *That cannot happen with a regular bank account. A person losing the PIN for her credit card (or her account number) can ask the bank to issue a new PIN after having proved her identity* (*e.g., showing her passport).*

So this is how bitcoins are stored. But how can we get some bitcoins? The most common way to get bitcoins is to buy them with some known currency like euros or dollars. There are a number of websites –called exchanges– created for that purpose.[13] Once you have bought bitcoins you tell the exchange to send them to your Bitcoin address. Et voilà! Another way to get bitcoins is to sell something and get paid in bitcoins. There is a third way to acquire bitcoins: to mine them.

2. Mining

In the Bitcoin system, mining is a twofold activity: processing transactions and creating new bitcoins. For each block added to the blockchain there was one miner who was the first to construct this block and send it to all the other miners with the message "please add this new block to the blockchain." One of the key aspects of Bitcoin is that there is a competition between miners to be the one constructing the next block. Each time a block is added to the blockchain new bitcoins are created and they constitute what is called the *block reward.* It is awarded to the miner who creates the block. That miner also collects all the transaction fees associated with the transactions in the block.

[12] http://fortune.com/2017/11/25/lost-bitcoins

[13] Some of the most popular websites are coinbase.com, bitstamp.com or blockchain.info, but these are not the only ones.

If there is a competition, it necessarily means that creating a block of transactions is not easy: it is not enough to simply create a list of transactions and send it to the other miners. In this section we explain what miners have to do to construct a block. Section III focuses on the competition –how it works and why it is a necessary element of Bitcoin system.

Sending bitcoins to a Bitcoin address consists of sending a message to the Bitcoin network through the internet.[14] With some simplification, we can say that such message contains:

- The sender's Bitcoin address;
- The recipient's Bitcoin address;
- The amount transferred;
- The fees the sender is paying to the miner who will process the transaction. The amount of the fee is decided by the sender (it can be zero);
- A "signature" by the sender.

The miners observe the transactions that have not been processed yet (*i.e.,* that are not in the blockchain) and they can choose which transactions they include in the block they will propose. This is where the fees come into play: a miner is more interested in processing transactions that carry higher fees.

For each transaction that has been selected, the miner, who has a copy of the blockchain, first checks whether the transaction is valid. To do that, the miner verifies whether the sender's address has the bitcoins it attempts to send by parsing the blockchain and checking that the address has received the bitcoins in the past but has not yet spent them. The miner also makes sure that the sender is indeed the owner of the sender's Bitcoin address. This is where the "signature" comes into play. The signature is generated using the private key and the message. Cryptographic tools are amazing: they permit verifying that the signature has been generated by the private key associated with the Bitcoin address without knowing what the private key is! In other words, the signature permits the miner to authenticate the sender, that is, to be sure that the sender is in possession of the private key associated to the senders' Bitcoin address (and thus likely to be the owner of the address).[15] The fact that the signature

[14] The Bitcoin network is simply a network of computers connected to each other through the internet.

[15] We say "likely to be the owner" because it could be a thief who stole someone's wallet.

is generated using not only the private key but also the message implies that the signature changes for each transaction. So any transaction re-utilizing a signature of a previous transaction will be immediately understood as being fraudulent and rejected by the miners.

Once the miner has verified the transactions and included them in her block, she will add a new special transaction that consists of awarding herself the block reward: newly created bitcoins (the amount of which is specified by the Bitcoin protocol). We almost have a block that is ready to be added to the blockchain. What is missing is a number, which is necessary to "match" the new block with the blockchain, a little bit like matching two pieces in a jigsaw: the new block must be "compatible" with the blockchain, for otherwise the other miners will refuse to add that block to their copies of the blockchain.

That number is a solution to a difficult numerical puzzle that cannot be solved by skill.[16] The only way is by trial-and-error: trying all possible numbers one after the other until we find the right one (we will see in the next section why it is difficult). That puzzle depends on the information in the current blockchain and in the (potential) new block. So knowing a solution for the previous block does not help in finding a solution for the next block. The puzzle is difficult, but once we have a solution, it is very easy to check that it is indeed the right number. In very, very simple terms, it is like searching for the square root of a very large number with a calculator that can only do multiplication. It takes time to manually find the square root of, say, 1,619,220,498,932,521, but it is very easy, however, to check that 40,239,539 is the solution. Similarly, while it is difficult to solve the Rubik's cube it is very easy to check whether it has been solved. The idea that a block of transactions can be added to the blockchain only after a miner has found a solution to a difficult puzzle is called *proof-of-work*.

Since a miner has no other option than to try many numbers until she finds the right one, finding the solution takes time. Bitcoin is designed so that, on average, it takes about 10 minutes for one of the miners in the Bitcoin network to find a solution. Sometimes it takes only a few seconds (the miner was lucky) and sometimes it can take 20 or 30 minutes. But the average is 10 minutes.

Once a miner has found a solution to the numerical puzzle, the block is ready to be sent to the other miners. If the miner is the first one to announce the creation of a new block, the other miners will add the block to their copy of the blockchain (after checking that the block contains authorized transactions

[16] For the interested reader: That puzle is built upon an algorithm called hashing, which is well known and widely used algorithm in cryptography. See https://en.wikipedia.org/wiki/Cryptographic_hash_function

and that the solution is correct). If at the beginning all miners have the same copy of the blockchain and if they all add the same block, then they all have the same new version of the blockchain.

An important aspect of mining is the following. Suppose that a miner, say, Alice, was still working on a block (*i.e.,* searching the solution of the puzzle) when a new block was announced and added to the blockchain. This means that Alice lost the competition, and she must start again from scratch –looking for a solution to the next new block. There are two reasons for that. First, the newly added block may contain some transactions that Alice was trying to process. So Alice would need to update the list of transactions she wants to process. Second, and more importantly, as we explained above, the puzzle that Alice has to solve depends on the blockchain. Since the blockchain has changed (a block has been added) the puzzle that Alice was working on is no longer the correct one. So Alice's efforts to find the solution to her old puzzle are wasted.

Each time a block is added to the blockchain the successful miner obtains a *block reward* in addition to the fees attached to the transactions she processed. This reward is made of new bitcoins, created ex-nihilo. This is the only source of new bitcoins. At the inception of the system, the reward was 50 bitcoins. By design, the reward is divided by two approximately every 4 years. In 2018 the reward is 12.5 bitcoins and it is expected to drop to 6.25 bitcoins around January 2020. Around May 2140 the reward will drop to 0. After this date no more bitcoins will be created and the only source of income for the miners will be the transaction fees, that is, the amount that the users pay to the miners for processing their transactions.

III. THE BITCOIN GAME

Mining consumes electricity and requires investment in computational power. Recently, both electricity consumption and the required investment increased significantly. Specialized mining computers cost several thousands dollars. And at the end of 2017, it was estimated that Bitcoin mining consumed as much energy as Denmark.[17] Why is it happening?

This happens because miners find it worthwhile to do so. Their incentives reflect the competitive nature of mining. Since only the first miner to find the solution gets rewarded new bitcoins and the fees, it is worth for a miner to invest in more computing power, to be quicker than others. This pushes the

[17] See, *e.g.,* https://digiconomist.net/bitcoin-energy-consumption; https://arstechnica.com/tech-policy/2017/12/bitcoins-insane-energy-consumption-explained/

other miners to also invest and be even quicker. This, by the very nature of the "tournament" (*i.e.,* only the winner gets the whole reward), leads to an arms race between the miners, which is exacerbated when the price of Bitcoin increases. By the end of 2017 one bitcoin was worth approximately $16,000, and thus the reward for mining was around $200,000. With such a high stake, the potential reward is worth investing in a powerful computer!

Thus, there are more miners, with more powerful machines, consuming even more energy. Interestingly, most of this effort, investment and energy consumption is wasted. Several thousands miners consume energy looking for a solution to the puzzle but only one will get to be the initiator of the next block. None of the calculations done by all other miners enters into the blockchain. If it was just for the sake of recording transactions on blockchain, it could be done with much less resource use. So, does it make sense, or is it a flaw in the system that needs to be fixed?

It turns out that in the Bitcoin system, this "wasted" computational effort plays an important role in the security of the system. All the calculations done by the "losing" miners increase the cost of winning the right to add new block to the blockchain, which helps prevent double-spending. To demonstrate this, imagine the following situation. Zoe buys a bike with her bitcoins. To do so, she sends some bitcoins to the seller and in exchange she gets the bike. Thus, the transfer of bitcoins from Zoe's address to the seller's address will show up in the blockchain. Suppose now that Zoe is not honest; she would like to get her bitcoins back –without giving the bike back to the seller. To do that, she would need to erase the transaction from the blockchain.[18] Changing this in her blockchain does not get her far in the Bitcoin design: the other miners only add blocks to their copies of the blockchain, they cannot delete or rewrite blocks. What Zoe would need to do instead is to persuade the other miners that they have a wrong copy of the blockchain and that the right copy is the one she constructed (which happens to be identical to the true one, except that it does not contain her transaction).

1. Forks

But wait, why would the other miners accept replacing one blockchain with another? This is by design. Bitcoin would not work if this was not possible.

[18] An alternative would be to replace the transaction with one where Zoe sends her bitcoins to herself (*e.g.,* to another address she owns) instead of the bike seller's address. The mechanics would be the same. However, she cannot create a transaction where the seller sends the bitcoins back to her. For that, she would need to know the private key of the seller, and she would also need to do it before the seller spends his bitcoins.

The reason is that sometimes having two (or more) different versions of the blockchains is inevitable. Here is why.

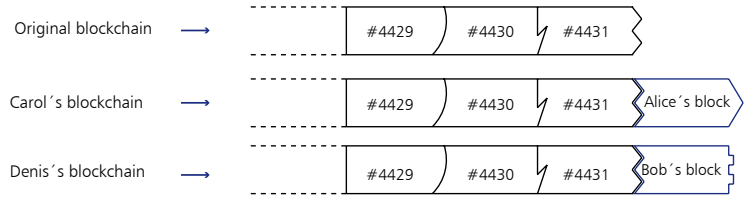
It can happen that two miners, say, Alice and Bob, find the solution for the block they are processing at roughly the same time. Also, since it is up to each miner to select which transactions to process, it is very likely that Alice's and Bob's blocks are not identical; that is, they do not contain the same transactions.

Since communications are not instantaneous on the internet, some miners will receive Alice's block before Bob's block while other miners will receive Bob's block before Alice's block.[19] What happens in this case? Consider the case of a miner, Carol, receiving Alice's block first. She will find that block valid and add it to her copy of the blockchain. Now, a few seconds later she receives Bob's block. Bob's solution was obtained for the blockchain without Alice's block. If we add Alice's block to the blockchain, Bob's block (with its solution to the numerical puzzle) is no longer compatible, like two pieces of a jigsaw what do not fit together. So Carol will refuse to add Bob's block to the blockchain.

At the same time another miner, Denis, has a symmetric problem. He received Bob's block first and added it to his copy of the blockchain. When Alice's block arrives a few second later Denis will find it invalid and reject it. So, now we have two competing copies of the blockchain: one with Alice's block and one with Bob's block. Figure 1 illustrates this situation.

FIGURE 1

TWO COMPETING BLOCKCHAINS



In the example depicted in Figure 1, the original blockchain finishes with the blocks numbered 4429, 4430 and 4431, with block #4431 being the last block added to the blockchain. We can see that the blocks fit together like in a jigsaw. The shapes of the blocks in the table capture the fact that the solution

[19] Transferring data over the internet can be fast, but it cannot be instantaneous because the speed of data transmission is bound by the speed of light. For instance, there are a bit more than 10,000km between New York and Tokyo, so it is *impossible* to take less than 67 milliseconds to send data from one city to the other. If you manage to transmit data faster than this, then you have proven Einstein wrong.

of the puzzle is what makes a block compatible with the blockchain (recall that the puzzle depends on both the blockchain and the block that is being processed).

Both Alice's and Bob's blocks can be added to the blockchain: The solution they found makes it possible to add their block to the original blockchain. However, once we have added Alice's block we cannot add the block proposed by Bob: it does not fit the new blockchain (Carol's). A similar situation happens if instead of adding Alice's block we add Bob's block. If we want to add Alice's block at the end of Denis's blockchain we need to find a new solution so that the left side of Alice's block fits with the right side of Bob's block.[20]

In the Bitcoin language when there are two or more competing blockchains we say that the blockchain *forked*. It happens about once a week. So, what happens next? Obviously, Alice will try to add a block to Carol's blockchain. She has no interest in Denis's blockchain because Alice's reward shows up in Carol's blockchain but does not in the version of the blockchain that Denis has. The same happens for Bob, who will work on Denis's blockchain. As for the other miners, they are indifferent. Some will work on Carol's blockchain and others on Denis's. Since the time needed to find the solution of the next block is never exactly 10 minutes, sooner or later one of the two versions of the blockchain (Carol's or Denis's) will be longer, *i.e.,* it will have more blocks. The convention in the Bitcoin system is that miners always focus on the longest blockchain. This ensures that in the long run there is consensus on which is the "true" blockchain. So the longer blockchain will be the winner and the other version of the blockchain will become orphaned.

Remark 3. *A transaction that appears in an orphaned version of the blockchain but not in the "winning" version of the blockchain is not lost. For the miners working on that blockchain that transaction is still in the pool of transactions that have not been processed yet. Because the blockchain can sometimes fork, most people wait a few blocks before considering that a transaction is indeed recorded in the blockchain. If I received bitcoins and that transaction is stored in the last block that has been added, most people will refuse my bitcoins if I try to spend them immediately. They will only accept my bitcoins if they show up in the 6th, or the 7th, or the 8th, … previous block (the convention is to wait at least 6 blocks, which is about 1 hour after a transaction has been processed). So, payments in bitcoins are not completely instantaneous.*

[20] Alice's and Bob's block may also have some transactions in common. If Bob wants to create a block after Alice's block he may then need to first reconstruct the set of transactions that he wants to process, eliminating the transactions already processed by Alice.

2. Rewriting History

Let us go back to our initial problem: Zoe bought a bike using some bitcoins and she would like to "erase" her payment. To do that she needs to:

1. *(Easy)* Locate the block where the transaction is;

2. *(Difficult)* Re-construct the block without her transaction (and thus solve the numerical puzzle again);

3. *(Extremely difficult)* Add blocks that came later (solving the numerical puzzle for each of them) and *be fast enough* so that the blockchain she constructs becomes longer than the original blockchain (and thus the other miners will switch to her blockchain).

Point 3 is what makes Bitcoin very difficult and costly to hack. During the time Zoe constructs blocks on her version of the blockchain the other miners do not stay idle! They continue to work on the main blockchain. What determines the probability that her attack succeeds is her share of computing power when compared to the total computing power across all Bitcoin miners. For instance, if there are 10,000 miners, each with the same mining equipment, then each miner has 0.01% of the computing power.

If Zoe has a small share of the total computational power it is extremely unlikely that she will be able to solve a number of puzzles faster than the rest of the miners. A higher share of computational power will increase the chances of success. And having 50% of the computational power ensures an attacker that she will eventually construct a longer competing blockchain.[21] But that much computational power is very expensive to acquire and operate: It would amount to paying the electricity bill of half of Denmark! Note that technically it is not impossible to forge or rewrite transactions. It is just unlikely and very expensive. Thus, potential attackers do not find it worthwhile.

Bitcoin's safety mechanism has two interesting properties that may seem counterintuitive. First, a larger number of "losing" miners who "waste" their computational effort make Bitcoin more secure against the attack. Second, when the Bitcoin price increases, it also makes the system more secure. To see the first point, compare the following two scenarios under which an attacker (*e.g.,* Zoe) is constructing a different version of the blockchain:

[21] Some detailed analyses show that around 30% or 40% is, in fact, enough to have very high probability of constructing a longer competing blockchain. See, *e.g.,* Kiayias *et al.* (2016).

- *Scenario A:* Besides the attacker, there are 10 honest miners (*i.e.,* miners working on the "true" blockchain). That is, there are only 11 miners on the Bitcoin network.

- *Scenario B:* Besides the attacker, there are 1,000 honest miners (so 1,001 miners in total).

For simplicity we assume that every miner in either scenario is equipped with an identical mining computer, *i.e.,* they all have the same nominal computational power.

Under either scenario *A* and *B* the honest version of the blockchain will grow at roughly the same pace, whether the attacker builds her own version of the blockchain or continues to work on the honest blockchain –on average one block every 10 minutes. This feature is embedded in Bitcoin's protocol. The average of 10 minutes is obtained by adjusting the difficulty of the puzzle. Without adjusting the difficulty of the puzzle a larger number of miners with more computational power would keep finding solutions to the puzzles, and mining new bitcoins, faster. The same happens when, for instance, you do not remember where your keys are. The more people searching for your keys, the faster you will find them. In order to keep the release of new bitcoins at a steady pace (on average), the Bitcoin protocol periodically adjusts the difficulty of mining. As a consequence, the difficulty of the puzzle depends on the *total* number of miners (*i.e.,* the total computational power involved in mining).

So what is the difference between scenarios *A* and *B* for our attacker? Since there is more computational power involved in mining in scenario *B,* the puzzle is much more difficult. This means that for the attacker, who is *alone* working on her version of the blockchain, finding solutions to the puzzles under scenario B will take much more time than under scenario *A.* And therefore, under scenario *B* she is *much less* likely to succeed in building a longer competing blockchain.

Under both scenarios, for each block there is only one miner who wins the competition, every 10 minutes or so. But what matters is that in scenario *A* there are only 9 "losers" who waste their computational effort, whereas in scenario *B* there are 999 losers, wasting much more computational effort. But the more "losers" there are, the more difficult the puzzle is, and thus the more difficult it is to attack, and hence the more secure Bitcoin is. This is why the "wasted" energy (by the miners who lost the competition) is a crucial element of the Bitcoin system.

This leads us to the second property of the Bitcoin's mechanism: when the price of Bitcoin increases, it becomes even more unlikely and increasingly

expensive to rewrite the blockchain history. A pricier Bitcoin means a more valuable mining reward, and thus more miners will find it worthwhile to invest in more computational power to participate in the competition to win bitcoins. This increases the total computational power involved in mining (and also the total energy consumption related to mining) and thus increases the difficulty of the puzzles. Now, that means that either the attacker's proportion of computational power decreases, which decreases the chances of attacker's success; or it forces the attacker to acquire and run more computational power to maintain her share of computing power, which is expensive.

The possibility of obtaining the block reward implies that there are always many other miners and adding a new block can only be done after intensive computations. This also makes creating alternative blockchains just as expensive, and thus makes rewriting history virtually impossible to be successful and worthwhile at the same time. So, Zoe keeps the bike and doesn't try to get her bitcoins back…

3. Security Versus Energy Waste

As we have seen, there are two types of forks: accidental and deliberate. Accidental forks are a natural part of Bitcoin's system. They are a consequence of the fact that Bitcoin is a distributed system, and the consensus mechanism that Bitcoin uses –proof-of-work– to assure consistency of this distributed ledger. When multiple blockchains occur accidentally, everyone wants the multiplicity to be resolved as soon as possible. The rule calling for following the longest blockchain serves this purpose, as it allows all miners to smoothly coordinate on one branch of the fork.

For deliberate forks, however, such coordination rule is not enough, as the attacker may feel tempted to create a longer blockchain with fraudulent transactions. The fact that forks can be innocuous may help the attacker to offer an alternative history without being detected as an attacker. Costly mining, however, makes creating such deliberate forks unlikely to succeed and very expensive, and therefore not worthwhile for the attacker.

So while mining consumes huge amounts of energy, and arguably most of it is "wasted", it should be considered as the price we need to pay for achieving security of a system without a trusted third party. In the Bitcoin system, it is not a flaw, but a crucial element that allows successful functioning. It is worth noting, though, that the computer science community is looking for alternative consensus mechanisms that would allow for similar security without a trusted third party, and without the wasteful mining. So far, however, none of these

alternative consensus mechanisms have proven to be as reliable as Bitcoin's proof-of-work.

IV. IS BITCOIN USEFUL AS A CURRENCY?

Bitcoin has been created with the intent to introduce a new currency. But is it really a useful currency? The crucial role of money is to serve as medium of exchange. There are many conditions that a useful medium of exchange needs to satisfy.[22] One of them is that we expect a good medium of exchange to keep its value between one transaction and the next one. If a "currency" is experiencing hyperinflation or high volatility of value, it is not really useful as a medium of exchange.

If a bike is worth $100 today but $1,000 tomorrow, and only $20 after tomorrow, trading will be less appealing. Buyers worry they would overpay and merchants worry that by the time they would use the money they receive its purchasing power will be significantly lower than expected. A solution would be to increase the price, but that would further repel the buyers. In most countries prices of goods and services fluctuate over time (usually due to inflation), but are relatively stable from one day to the next. What about bitcoins? Figure 2 shows a typical day for Bitcoin. The difference between the lowest price that day (about $14,336) and the highest (about $18,353) is roughly 28%! There are, of course, many days where the price of Bitcoin does not change as dramatically as in Figure 2, but there is a large consensus to say that the price is very unstable. This may discourage many people from using it as medium of exchange.

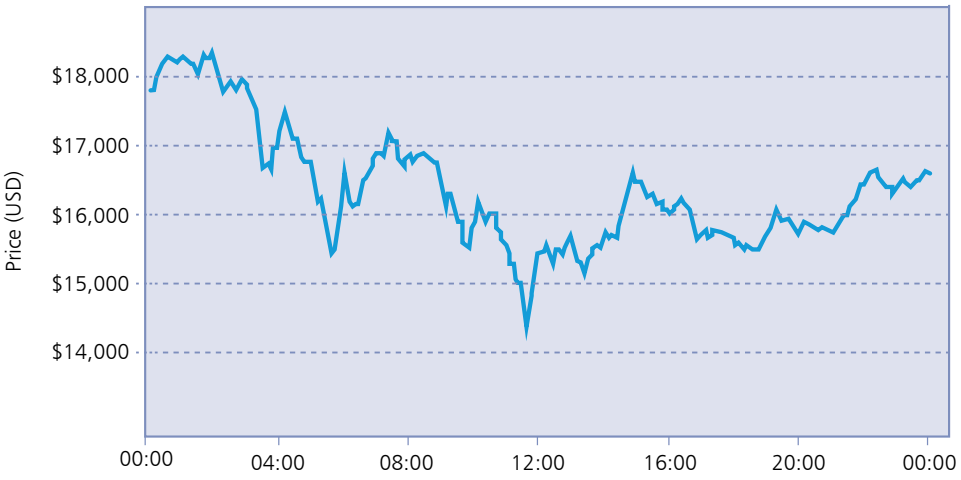
Another reason is that, for now, using bitcoins is not as easy as using a credit card or payment systems that some cell phones are equipped with. Moreover, potential users often indicate that existing methods of payment serve their needs well, and they see no reason to adopt a new one (see, *e.g.*, Survey results in Henry, Huynh and Nicholls, 2018).

There is also a technical reason that may limit wide-spread adoption of Bitcoin. Recall that Bitcoin transactions are processed in blocks, at a rate of one block every 10 minutes. By design a block in the blockchain cannot exceed certain size, which severely limits the number of transactions that can be processed by the Bitcoin network: at most 7 transactions per second. In comparison, VISA

[22] Economists often cite a well-worn definition that money is a medium or exchange, unit of account, and store of value. The latter two are simply necessary to fulfill the first role well. But there are other conditions, *e.g.,* we need money to be reasonably divisible. In the context of physical tokens, useful money also needs to be uniform, durable, portable, etc.

FIGURE 2

PRICE OF BITCOIN ON DEC 8, 2017



Source: coinmarketcap.com.

handles on average a few thousands transactions per second and can handle up to 56,000 transactions per second (which is almost 5 billions per day!).[23]

In 2017, on many occasions, the demand for Bitcoin transactions exceeded the system's capacity. In such cases, there may be a significant delay in processing transactions. Users may ensure their transactions are processed and included in the blockchain more quickly by offering higher fees to the miners. And so, towards the end of 2017 Bitcoin fees increased multiple times to as much as $30 per transaction.[24] For most transactions it is higher than the credit card fee!

Nonetheless, in some cases people may prefer to use Bitcoin for trading rather than available alternatives. For them the benefit that only Bitcoin offers outweighs the risk of value volatility, inconvenience of interface, and the cost of higher fees. Illegal traffic (drugs, arms), gambling or tax evasion were the first areas where we saw a more frequent use of bitcoins. The relative anonymity of Bitcoin and its speed made it attractive for the people involved in such activities.

There are some reported niches of legal activity that use Bitcoin, such as private aviation. There, relative privacy combined with large value transactions

[23] https://usa.visa.com/dam/VCOM/download/corporate/media/visa-fact-sheet-Jun2015.pdf

[24] See, *e.g.,* https://news.bitcoin.com/miami-bitcoin-conference-stops-accepting-bitcoin-due-to-fees-and-congestion/

make Bitcoin appealing both to the merchants and to the buyers. Buyers gain more privacy and are not constrained by credit limits on cards. Merchants receive money quicker than via credit cards or wire transfers, and save on credit card fees.[25] Notice that for transactions amounting to tens of thousands of dollars, credit card fees may exceed the $30; not so for merchants making large numbers of small ticket transactions. This asymmetry comes from an interesting fact that credit card fees are based on the value of transactions, but Bitcoin fees are based only on the urgency for the user.

Relative privacy and the decentralized nature of Bitcoin provide useful properties for people living in countries with high currency controls or risk of government interference with bank accounts, *e.g.,* when the government limits the daily or weekly withdrawals an individual can make, or when the government can confiscate part of the holdings. Also, despite high volatility, Bitcoin may be preferred to currencies with very high inflation. For people living in countries like Venezuela or Zimbabwe, Bitcoin can be very attractive.

Nonetheless, despite the few niches, 8 years on, Bitcoin has not become a popular medium of exchange. One could think that the story of Bitcoin stops here. It does not.

V. WHAT'S NEXT?

The "Bitcoin revolution" did not stop with the creation of Bitcoin. It has just started. Bitcoin opened the door to a vast array of innovations and ideas, which can be divided into two categories, cryptocurrencies and non-currency applications.

1. Improving on Bitcoin: Competing Cryptocurrencies

When we talk about cryptocurrencies almost everybody thinks about Bitcoin, and many people think that this is the only cryptocurrency. But it is not. By the end of 2017 there were more than a thousand different cryptocurrencies traded on digital cryptocurrency exchanges, that is, these cryptocurrencies could be bought or sold for other currencies such as the US dollar or the euro (although many of them are only tradable with bitcoins). Most of these currencies have a very low price, so low that the total market capitalization (*i.e.,* the worth of the total number of coins in circulation) is only a few thousand US dollars.

[25] It takes only a few hours to send bitcoins to the other side of the planet while it takes several days with international bank transfers. Services like Western Union are fast but the fees they charge can be higher and for some countries they do not offer as much anonymity as Bitcoin for large amounts.

In comparison, Bitcoin's market capitalization on December 31st, 2017 was around $240 billion! At this date there were just under 40 cryptocurrencies with a market capitalization above $1 billion. The top cryptocurrencies in terms of capitalization aside from Bitcoin are Ripple, Ethereum and Bitcoin Cash. Like Bitcoin, these cryptocurrencies are also very volatile. For instance, Litecoin, another cryptocurrency, had a capitalization around $20 billion on December 19th, 2017, but it dropped to nearly $14 billion two weeks later.

Why are there so many cryptocurrencies? There are many reasons. One of them may be obvious: the hope to make a quick buck. Given the rapid growth of Bitcoin some people expect that they can raise a lot of money if they create or spot the "next Bitcoin." This is why so many of the cryptocurrencies were really just a copy of Bitcoin. Bitcoin is an open source project. Anyone can make use of it, with our without changes. Anyone can put a Bitcoin-copy-cat on the market. Not surprisingly, however, most of these copy-cats were not successful, and aside from some pump-and-dump speculative schemes, they reached only low market capitalizations.

The possibility of alterning the open-source code of Bitcoin, however, gave raise to cryptocurrencies developed for a different reason: improvement on Bitcoin.

As we have explained in Section IV, Bitcoin has a number of shortcomings. One is that by design, it can process at most 7 transactions per second. Therefore, should a cryptocurrency really become a popularly used means of payment, it cannot be the Bitcoin as we have it now. There are two possible ways to fix that. The first possibility is that Bitcoin evolves and its software is updated in order to allow for a larger flow of transactions. The problem with that is exactly where Bitcoin's innovation lays; it is too decentralized. There is no governance structure where some authority or committee would enforce changes in the Bitcoin protocol. Any modification to the Bitcoin protocol requires an overwhelming majority of the miners, if not unanimity. This is very difficult to achieve. A group of miners and Bitcoin users proposed a change in the summer of 2017. It needed 95% of miners accepting the change to be adopted. Such a strong majority was not achieved. Thus, the Bitcoin protocol was not changed. However, a significant portion of miners implemented the proposed change, which started a new version of Bitcoin that they called *Bitcoin Cash.* It exists in parallel to Bitcoin. And this parallel existence points to the second way in which a cryptocurrency could become a popularly used means of payment; another, improved cryptocurrency that can handle larger volumes of transactions would surpass Bitcoin in adoption and dominate the market.

Another issue with Bitcoin is that the total amount of coins that will ever be mined is fixed. In the long run this will create deflationary dynamics. That

is, instead of having prices going up (inflation), prices decline. A one time drop in prices like a Black Friday is a blessing for consumers, but it is not if the decline is persistent. To demonstrate this, imagine you have a mortgage with monthly payments of, say $1,000. If prices drop for a long period of time so will your salary, but not your monthly payment. This means that the share of your income that goes to your mortgage keeps increasing. The same applies to firms that took a loan for their investments. Another effect of deflation is that consumers wait to purchase durable goods (like a washing machine), especially larger ticket items, as long as possible. If everyone thinks the same way, this stifles the economic activity.

Many of the cryptocurrencies that were created after Bitcoin tend to correct one or more of the "flaws" of Bitcoin. For instance, Litecoin is a cryptocurrency that aims to process transactions faster than Bitcoin. With Litecoin, the average time to add a block to the Litecoin blockchain is only 2.5 minutes (versus 10 minutes for Bitcoin), and blocks are designed so that Litecoin can handle up to 56 transactions per second (only 7 per second for Bitcoin). Litecoin uses also different algorithms for the puzzle miners have to solve, which originally aimed to reduce the amount of energy used in the mining process (see, *e.g.,* Gandal and Halaburda, 2016). Many of the existing (and forthcoming!) cryptocurrencies follow the same motivation behind the creation of Litecoin: improve on Bitcoin's design. And despite the currently overwhelming popularity of Bitcoin, it is possible that the cryptocurrency of the future is one of the competing designs.

2. Smart Contracts: The Ethereum and Beyond

A major new development in the area of cryptocurrencies and blockchain is the creation of the *Ethereum platform.* Ethereum was proposed in late 2013 by a young programmer, Vitalik Buterin. The development of Ethereum started soon after (financed through a crowd-sale in the summer of 2014) and was officially launched on July 30th, 2015.

The Ethereum system is similar to Bitcoin in many aspects: it has its blockchain and miners, and a cryptocurrency, called *ether*. It offers new functionality above Bitcoin and other cryptocurrencies by focusing on smart contracts.

Smart contract is a set of instructions that are automatically performed if some conditions, set by the user, are satisfied. A very simple example of a smart contract would be the following. Consider two users, say Alice and Bob, where Alice is selling her house to Bob. Bob can pay Alice through the Ethereum network, but he adds in his transaction the following smart contract: Alice receives the ethers corresponding to the transaction only if, before a certain

date, the property registry (maintained by the local authorities) indicates that the house now belongs to Bob. In this example the smart contract would work as follows. The program (*i.e.,* the smart contract) periodically checks if the property registry indicates Bob as the owner. If it is the case then Alice receives the ethers Bob sent as a payment. Of course, such a smart contract is only possible if the property registry can be accessed remotely by a program.

Another potential use for smart contracts is the payment of demurrage in the shipping industry. When a container arrives at some destination its arrival is registered by the port authorities in a database. This can trigger the execution of a smart contract that finalizes the payment between the seller and the buyer, which would include the penalty paid by the seller (or the shipping company) if the container arrives late.

Smart contract may look like a fancy phrase but the concept was not born with Ethereum. The automatic payments we set up to pay our utility bills, rent or mortgage are, in fact, smart contracts. The advantage of Ethereum's smart contracts, compared to those offered by our banks, is that they can be designed any way we want. Bitcoin allowed rudimentary smart contracts to be stored on blockchain. But it was only Ethereum that allowed any two parties to include in its blockchain any smart contract that could be programmed. This allows for greater flexibility and functionality.

Of course, the attractiveness of smart contracts is predicated on the ability to independently confirm with non-related parties that some conditions have been satisfied. In the above example with Alice and Bob, for the smart contracts to work, property deeds need to be digital and open to programs that want to parse them. For shipments we also need port authorities and transporters to record their actions in databases that are accessible remotely by third parties. Smart contracts like the ones we described are not yet ubiquitous because the infrastructure needed for them is still nascent (trackers for shipment, legal databases accessible remotely, etc).

Fundamentally, those smart contracts do not need a blockchain solution like Ethereum: a centralized database would also work. However, the possibility of creating custom smart contracts for many people is a game changer that could result in significant cost savings.

3. Alternative Functionality of Cryptocurrencies: Initial Coin Offerings

Market activity as well as academic analysis indicate that many people are acquiring Bitcoin and other cryptocurrencies solely as an investment, not because

they expect them to be widely used cryptocurrencies. Nothing demonstrates the investment potential of cryptocurrencies better than the proliferation of ICOs, *initial coin offerings,* since 2016.

Not accidentally, "ICO" sounds like IPO (initial public offering). IPO is a way for a company to raise money and increase its capital by selling stakes in the company, in the form of shares, to the public. After the initial sale, the shares are traded on stock exchanges like Nasdaq. Regulators put significant requirements on the company before it is allowed to sell its shares to the public in order to make sure it is not going to collect the money and disappear, rendering the shares worthless. Thus, traditionally, only companies that are sufficiently large and stable have been able to bear this regulatory burden and have an IPO. Also, becoming public is often seen as an achievement for a company. But due to its requirements, preparing for an IPO is costly, time consuming and risky. Not all companies starting the process are eventually authorized to launch an IPO.

Cryptocurrencies (the "coins") became an attractive alternative for some companies, dispensing with the paperwork and requirements imposed on a company going public. An ICO is in fact like crowdfunding, and in this way similar to an IPO. Someone proposes a project and individuals or investors are asked to contribute. In the case of an IPO, contributors obtain shares in the company. For crowdfunding, the typical case consists of obtaining an object at some discount or being among the first ones to get it. In the case of an ICO contributors obtain *tokens*. It is not clear, however, what a *token* represents.

Sometimes the ICO is for the launch of a new cryptocurrency. In this case the *tokens* represent the new coins. That is, a contributor to the project is awarded *tokens* that are, at a later date, converted into coins of the new cryptocurrency. This is what happened for instance with Ethereum.

It is tempting to view the ICO *tokens* received by a contributor as shares of the company. But in most cases they are not. There are no rules or oversight on what the tokens represent or what value they offer to the owners. Investing in an ICO is thus very risky because there are no contractual guarantees that an investor will eventually receive shares in the company, let alone dividends.

4. Looking Forward: Blockchain Applications

For many people the innovation brought by Bitcoin does not only consist of a decentralized cash system. They see potential in the concept of a blockchain –and how it is maintained– for purposes other than a cryptocurrency. After all, it seems, a blockchain is merely a database.

Technically speaking Bitcoin's blockchain is a distributed and permissionless database. What does that mean? *Distributed* means that multiple network members can make changes to the database. The challenge in such a case is to ensure the consistency of such a database across different network members. Distributed databases have been used and researched for three decades. However, all distributed database designs before Bitcoin involved some third party who was managing access by the network members to the database, and often also served as an arbiter in case of entry conflicts; they were permissioned distributed databases. Bitcoin, in contrast, is a permissionless distributed database. Permissionless means that anybody can modify it (*e.g.,* become a miner in the case of Bitcoin): no permission from a third party is needed to do so.[26]

Satoshi Nakamoto did not invent the concept of distributed and permisisionless database. The cryptography community has been working on creating such a database since mid-1980's. There were several less successful attempts. Satoshi Nakamoto's invention in Bitcoin's design was to look beyond cryptographic solutions and also heavily rely on an economic incentive system to achieve the database's consistency. This led to an incentive scheme for the miners involving a tournament structure based on proof-of-work and a mining reward. A crucial future necessary for achieving consistency in this database is that it has an *append-only* structure, *i.e.,* we can only add new records into the database (which must be compatible with already existing records), thereby creating the so-called blockchain. Since the database was not designed to allow rewriting past entries, for any forking attempt the attacker has to create an alternative database (*i.e.,* an alternative blockchain), and provide conditions under which the alternative one would be accepted instead of the original database (*i.e.,* make it a longer blockchain). A drawback is that this append-only structure comes at the cost of speed when consulting the database. Most of the databases that governments, our banks and other corporations are using are more complex, but faster to consult (called *relational database*).

It did not take long before people envisioned that the "blockchain technology" could have possibly other uses besides maintaining a digital currency. Implementing a reliable permissionless distributed database could permit to dispense with third parties that are here to check and verify the veracity of the data. For instance, if we want to buy a house we will have to check that the seller is indeed the owner of the house. Similarly the seller will want to check that we have sufficient funds to buy it. All those operations involve lawyers, brokers and/or notaries. Many people believe that a blockchain technology (that would make all the necessary data for such transactions accessible) would allow us to dispense with those intermediaries. In the finance industry, banks

[26] Being able to add blocks to the blockchain only after having solved a difficult puzzle is not incompatible with being permissionless because anyone can become a miner.

and investors believe that if trades of securities (stocks, bonds, derivatives, etc.) are recorded on a distributed database it will be easier to track ownership and ease settlements (the transfer of the asset from the seller to the buyer).

Potential applications of the "blockchain technology" do not follow the exact model of Bitcoin, though. Many blockchain applications under development consider *private* (not everybody can download and scrutinize the data) and *permissioned* database (ability to modify or enter new data is granted by a third party). There is no established definition of a *blockchain*. It seems that the features that most agree on are *distributed and append-only.* This is departing from the Bitcoin's innovation that allowed for a distributed and *permissionless* database.

An example of such wide interpretation of the blockchain technology is the e-citizenship implemented in Estonia. In Estonia, voting, real estate, taxes, banking, relations with schools, or managing health care data are now done through the internet. The data is not stored centrally, it is stored in thousands of servers and the technology behind the platform (called X-Road) claims inspiration from the blockchain design.[27] Note that in this case we are stepping away from the "Bitcoin philosophy" because the Estonian's "blockchain" is ran by the government.

VI. CONCLUSION

Almost ten years after its inception Bitcoin managed to become a household name. Every day thousands of people buy or sell bitcoins and other cryptocurrencies, and some of them are even using bitcoins for what it was intended, a currency to buy or sell goods and services. There is now little doubt that Bitcoin's design is a succesful attempt to create a *decentralized* digital cash system. Whether bitcoins can claim the status of a currency is still an open question. Because of a number of Bitcoin's limitations, it is not adequate for mass use. But perhaps other cryptocurrencies will be.

Bitcoin's contribution is more than that of a cryptocurrency. The excitement about the properties of the blockchain that Bitcoin's technological design generated has opened up a discussion about its use for a wide range of applications. In effect, it has turned attention and prompted some innovative uses of smart contracts and distributed databases. If we think carefully, most developments and applications proposed are not new. The concepts of distributed database and smart contracts existed well before Bitcoin or Ethereum.

[27] The whole system is backed up on servers in Luxembourg in case of failure or an invasion by Russia.

For aficionados, blockchain based solutions are infinite. Current projects claiming to be based on blockchain range from voting *(Soverign)* or equity management (*Chain,* launched by Nasdaq) to internet domain registry *(Namecoin)* or file storage *(Storj).* But the truth is, as of today there is still no "killer application" for the technology.

So, is Bitcoin a revolution after all? We may say it is. Or we may argue and say that there is not much innovation around Bitcoin (*i.e.,* most of the concepts already existed). But there is little doubt that popularity around Bitcoin and blockchain has spurred the debate and the development of decentralized and automated solutions.

BIBLIOGRAPHY

Androulaki, E.; Karame, G. O.; Roeschlin, M.; Scherer, T., and S. Capkun (2013), "Evaluating user privacy in Bitcoin," in *International Conference on Financial Cryptography and Data Security,* Springer, 34–51.

Ferguson, N. (2008), *The ascent of money: A financial history of the world,* Penguin.

Gandal, N., and H. Halaburda (2016), "Can we predict the winner in a market with network effects? Competition in cryptocurrency market," *Games,* 7: 16.

Halaburda, H., and M. Sarvary (2016), *Beyond Bitcoin. The economics of digital currencies,* Springer.

Henry, C.; Huynh, S., K. P., and G. Nicholls (2018), "Bitcoin awareness and usage in Canada," *Journal of Digital Banking* (forthcoming).

Kiayias, A.; Koutsoupias, E.; Kyropoulou, M., and Y. Tselekounis (2016), "Blockchain mining games," in *Proceedings of the 2016 ACM Conference on Economics and Computation,* ACM: 365–382.

Nakamoto, S. (2008), *Bitcoin: A peer-to-peer electronic cash system*.

Radford, R. A. (1945), "The economic organisation of a POW camp," *Economica,* 12: 189–201.

Roberts, J. J., and N. Rapp (2017), "Exclusive: Nearly 4 million bitcoins lost forever," *Fortune,* November 25, http://fortune.com/2017/11/25/lost–bitcoins/

Urquhart, A. (2016), "The inefficiency of Bitcoin," *Economics Letters,* 148: 80–82.

BIG DATA AND COMPETITION POLICY

Adina CLAICI[1]

Abstract

This chapter reviews the main issues raised by competition policy enforcement in relation to big data. The exponential increase in the amount of data available in our society as well as the unprecedented development of technologies to collect, process, store and use that data spurred numerous questions in many areas. We focus on the competition field and aim at providing a balanced overview of different opinions regarding the degree of market power conferred by big data. Due to the multitude of business models adopted by different data-driven firms and the circumstances in which data is used, it is not possible to drive general conclusions as regards the characteristics that make big data more or less valuable as an asset. On the one hand, some academics and competition enforcers have identified certain theories of harm and potential anti-competitive effects stemming from the use of big data. On the other hand, big data and the associated technologies provide consumers with new products and innovative services. In essence, this debate is just a new version of the old question on how to protect competition without stifling innovation. Finally, we present a few landmark cases that have certain, albeit still limited relevance in the discussion about big data.

Key words: Big Data, competition, innovation.

JEL Classification: L80.

[1] The views expressed are purely those of the writer and may not in any circumstances be regarded as stating an official position of the European Commission.

I. INTRODUCTION

The amount of personal data spreading throughout the economy has increased exponentially in the past decade and continues to do so. This trend is intertwined with the development of machine learning techniques to explore, analyse, and use data as well as with new possibilities to store huge amounts of data. Big data provides the "raw material" for machine learning. For the purpose of the current chapter, we refer to machine learning as the field of computer science dealing with algorithms that allow "machines" (*i.e.,* computers) to "learn" and progressively improve performance on a specific task as more data is fed into them.[2] Algorithms are not new, but never before the discussion around them has moved so far beyond computer science to reach domains like economics, law, ethics or consumer protection. This chapter will focus on the questions raised by big data and machine learning in the area of competition policy.

It is uncontested that consumers and society have benefitted in an unprecedented way from this progress. *The Economist* calls data the fuel of the future, comparing it to oil, which was the driver of growth and change in the last century.[3] OECD describes possible efficiencies brought about by algorithms recognizing that "data-driven marketplaces are generally associated with significant efficiencies both on the supply and demand side" (see OECD, 2017). Algorithms may help improving existing products and services or developing new ones. They may also support consumer decisions by providing structured information that can be accessed quicker and more effectively and also by providing information on new dimensions of competition other than prices, such as quality or other consumers' preferences.[4]

Marr (2016) provides a comprehensive overview of how companies and organizations, big and small alike, across different industries, are using big data to deliver value in diverse areas. For example, supermarkets sell millions of products to millions of people every day. Having the right products in the right place at the right time, so that the right people can buy them, presents huge logistical problems. Timely analysis of real time data is seen as key to driving business performance. Online retailers, such as Amazon, rely heavily on data for making good predictions in order to minimize waste. Too much or too little stock would mean huge costs for the company. Another example is the streaming movie and TV service Netflix which accounts for one-third of peak-time Internet traffic in the US and collects and monitors data from millions

[2] A more comprehensive discussion around machine learning is provided in a different chapter of this book (Hansen, 2018).

[3] *The Economist,* May 6th 2017.

[4] As, for example, in the case of reviews and ratings.

of subscribers in an attempt to understand the viewing habits of customers. The business is built around using data and analytical techniques to predict what people enjoy watching. Besides describing well-known success stories, Marr (2016) recounts numerous examples of non-profit organization and small businesses that developed considerably by making use of data, algorithms and novel ideas on how to use them.[5]

However, this shift towards a data-driven economy is not coming without concerns in numerous directions. As the Competition Commissioner Margrethe Vestager said: "big data has enormous potential. But it won't achieve that potential unless people are confident that their rights are protected […] the future of big data is not just about technology. It is about things like data protection, consumer rights and competition. Things that give people confidence that big data won't harm them."[6]

This chapter remains in the realm of competition policy and aims at providing an overview of the *status quo* of the main issues raised by competition enforcement in relation to big data. Competition agencies around the world have started to look more systematically into potential benefits and harm stemming from big data and the use of algorithms. In 2016, the French and German competition authorities published a joint document on Competition Law and Data (Autorite de la Concurrence, 2016). Their paper identified some of the key issues and parameters that may need to be considered when assessing the interplay between data, market power and competition law. They put forward various theories of harm usually associated with data collection and exploitation in digital markets and discussed some of the parameters that are to be considered in assessing the relevance and credibility of these theories of harm.

Also in 2016, the FTC issued a Report called "Big Data: A Tool for Inclusion or Exclusion?" (FTC, 2016). Their study is intended to educate businesses on important laws and research ideas that are relevant to big data analytics and provide suggestions aimed at maximizing the benefits and minimizing its risks.

[5] One of the examples refers to a small local butcher in London, Pendleton & Son. When hit by competition from a supermarket chain, the local shop could not remain competitive in price hence decided to make use of the data to improve the product and service. Using sensors, the firm managed to measure how many people walked by the shop, how many stopped to look and how many came into the store. With the help of this information, the butcher was able to refine the display and to find out that a significant flow of people was passing by at late hours, due to other activities in the area. By adjusting the opening hours and offering products that those people required (Google Trends was helpful in finding this information) the butcher streamlined the activity and managed to become profitable again.

[6] See: https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/big-data-and-competition_en

Some of the first scholars who focused the discussion around the digital economy were Erzachi and Stucke (2016) and Stucke and Grunes (2016). They recognised the numerous appealing features of the online data-drives competition, such as: increasing market transparency and the flow of information, lower search costs, easier entry and expansion, more dynamic disruption and efficiencies, and overall empowerment of the consumers. However, after acknowledging that consumers reap many benefits from online purchasing, the message of their book is that the computer programmes and data-crunching that make browsing so convenient are also changing the nature of market competition, and not always for the better. They identified how the rise of sophisticated algorithms and the new market reality can significantly change our paradigm of competition for the worse –with more durable forms of collusion, more sophisticated forms of price discrimination, and instead of data-drive monopolies that, by controlling key platforms, dictate the flow of your personal data.

Maurice Stucke has also co-authored another book, *i.e.,* Stucke and Grunes (2016) providing an overview of potential theories of harm related to big data, including also a critical view on how some competition authorities have addressed data related issues in a few cases so far. They believe that traditional instruments of competition policy have to be adjusted to the new realities of the data-intensive companies. J. Kennedy (2017) examines and rebuts the theories put forward by Stucke and Grunes (2016) and expresses that data-rich companies are not a threat, but rather an important source of innovation, which policy makers should encourage, not limit. Moreover, he claims that competition policy has the right tools to deal with potential theories of harm stemming in relation to data.

The aim of this chapter is to gather views from academics, practitioners, industries and competition agencies in order to provide a balanced assessment of the potential pro-competitive and anti-competitive sides of the data-driven economy.[7] After a comprehensive introduction into big data and its characteristics, we point to possible challenges brought about by the digital economy in different competition policy areas, namely cartels, abuse of dominant position and mergers.

The chapter is structured as follows. Section two provides a brief description of the main characteristics of 'big data". The third section describes how data as an asset can confer market power to firms. Section four deals with collusion and enquires whether the increased amount of data may give rise to

[7] For the purpose of this chapter, the concepts of data-driven economy and digital economy are equivalent and refer to both 'big data' as well as the technology and intelligence to process it.

enhanced cartel activity. Section five puts forward potential theories concerning alleged abuses of dominant position through the ownership or use of data and discusses a few abuse of dominance cases in the sphere of tech giants. The sixth section briefly explores the area of mergers where data is an important asset and, finally, section seven concludes.

II. WHAT IS "BIG DATA"?

There is no universal definition of big data and there is no clear threshold when a certain amount of data becomes "big". Various practitioners tried to define big data though its main characteristics, which are often referred to as the four or five 'V's.[8]

- Volume of data refers to the vast amounts of data generated every second. The volume of data collected has increased significantly and will likely continue to grow. The costs to collect, store, process and analyse data have all decreased.

- Velocity refers to the speed at which new data is generated and the speed at which data moves around. Data is accessed, processed, and analysed much faster. This represents the time-value of data.[9]

- Variety of data refers to the multiple types of data firms now collect and use.[10]

- Value of data comes from the ties to big analytics (technical means to extract insights from the data). The volume, velocity and variety of data together with the algorithms to process it enable value extraction.

- Veracity refers to the messiness of the data. With many forms of big data, quality and accuracy are less controllable but big data and analytics technology allows working with these type of data.

Big data can be personal and non-personal. OECD (2013a) defines "personal data" as "any information relating to an identified or identifiable individual (data subject)". By contrast, non-personal data refers to data that is

[8] See for example: Stucke and Grunes (2016) or Marr (2014) available at: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/

[9] For example, geo-localisation technologies can recommend restaurants in real time.

[10] For example retailers create individual shoppers' profiles and can tailor better their offers.

not linked to a certain individual, such as, for example, weather data or traffic information.

Most controversy in both competition enforcement as well as consumer protection fields lies around personal data, as this is considered an asset linked to an individual, which is, at the same time, a consumer. OECD (2013a) lists, by way of example, the following types of personal data:

- User generated content, including blogs and commentary, photos and videos, etc.
- Activity or behavioural data, including what people search for and look at on the Internet, what people buy online, how much and how they pay, etc.
- Social data, including contacts and friends on social networking sites.
- Locational data, including residential addresses, GPS and geo-location (*e.g.,* from cellular mobile phones), IP address, etc.
- Demographic data, including age, gender, race, income, sexual preferences, political affiliation, etc.
- Identifying data of an official nature, including name, financial information and account numbers, health information, national health or social security numbers, police records, etc.

Others have further categorized personal data in different ways. For example, Schneier (2010) has developed a taxonomy of personal data, using social networking sites as an example, and differentiates the following types:

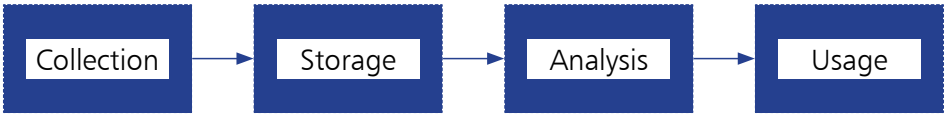
- Service data, which is the data you need to give to a social networking site in order to use it. It might include legal name, age, and credit card number.
- Disclosed data is what you post on your own pages: blog entries, photographs, messages, or comments.
- Entrusted data is what you post on other people's pages. It's basically the same as disclosed data, but the difference is that you do not have control over the data –someone else does.

- Incidental data is data the other people post about you. The difference from the disclosed data is that you don't have control over it and you did not create it in the first place.

- Behavioral data is the data that the site collects about your habits by recording what you do and who you do it with.

At its core, big data is about predictions, as Mayer-Schonberger (2013) describes it in his book. "It is about applying math to huge quantities of data in order to infer probabilities, such as the likelihood that an email message is spam or that the trajectory and velocity of a person walking mean he'll make it across the street in time –the self-driving car need only slow slightly." Furthermore, Mayer-Schonberger explains how the use of big data involves a shift in our problem-solving approach from causality to patterns and correlations. As humans we have been conditioned to look for causes whereas in a big-data world we can discover patterns and correlations in the data that offer us novel and invaluable insight. He argues that correlations may not tell us precisely why something is happening, but they alert us that it is happening. The author believes that in many situations this may be good enough and provides some interesting examples. For instance, if millions of electronic medical records reveal that cancer sufferers who take a certain combination of aspirin and orange juice see their disease go into remission, then the exact cause of the improvement in health may be less important than the fact that they lived. Likewise, if we can save money by knowing the best time to buy a plane ticket without understanding the method behind the airfare madness, that is good enough. Big data is about what, not why. We don't always need to know the cause of the phenomenon; rather, we can let data speak for itself (Mayer-Schonberger, 2013).

Finally, before discussing the competitive effects of big data, we will briefly describe the different blocks in the data value chain. It is important to make the distinction between the four stages in the value chain as the competition for data may vary across them and the barriers to enter also. Rubinfeld and Gal (2016) provide a simple, clear and intuitive characterization, as show in the diagram below:



Collection refers to the extraction of the data, storage to the load of data into powerful "deposits" and organisation into databases, analysis relates to the data processing in order to find correlations and usage is the final stage of extracting information relevant to making decisions.

III. BIG DATA AND MARKET POWER

This section attempts to address three issues. In the first place, we question whether big data can be considered valuable. Secondly, to the extent that it is, we enquire whether big data can create market power and barriers to entry. Finally, we briefly discuss the theory of big data as essential facility and the possibility to regulate data sharing.

1. The Value of Big Data

Big data represents a core economic asset that can create significant competitive advantage for firms and drive innovation and growth (OECD, 2013). OECD identified five sectors in which the use of data can stimulate innovation and productivity growth: online advertisement, health care, utilities, logistics and transport, and public administration. Overall, the benefits that big data can create in these sectors include: the development of new data-based goods and services; improved production or delivery processes; improved marketing (by providing targeted advertisements and personalised recommendations); new organisational and management approaches, or significantly improved decision-making within existing practices; and enhanced research and development.

In a subsequent paper (OECD, 2013b), OECD provided a survey of methodologies for measuring and estimating the value of personal data from a purely monetary perspective (*i.e.,* without taking into account the indirect impacts of the use of personal data on the economy or society). It looks at a range of measurement and estimation techniques and identifies the main benefits and drawbacks of each approach. The use of multiple methodologies helps reducing context dependent biases. The OECD report details at least six methods that can be used to measure the value of data:

- The most direct way to approach the value of personal data is to evaluate the market prices at which personal data are offered and sold. The values provide a market-based measurement based on supply and demand.[11]

- Another methodology could be based on market capitalisation of personal data, *i.e.,* the company value per user. This however leads to

[11] In 2013 when the OECD report was written, examples of prices in the United States for personal data ranged from USD 0.50 for a street address, USD 2 for a date of birth, USD 8 for a social security number, USD 3 for a driver's license number and USD 35 for a military record. These are only estimates but provide some insight into the relative market values of different pieces of personal data.

valuations that can fluctuate considerably, largely influenced by other economic factors over the period.[12]

- A third method similar to market capitalization per user consists in revenues or net income per record/user.[13]

- Fourth, the monetary value of personal data could be estimated via an assessment of the economic costs of a data breach.[14]

- Fifth, economic experiments and surveys could provide a range of prices that firms would need to pay individuals to give up some of their personal information. Even though research in this area is still in a preliminary stage, OECD could extract two general messages. First, people tend to differ with respect to their individual valuation of personal data (*i.e.,* the amount of money sufficient for them to give away personal data) and their individual valuation of privacy (*i.e.,* the amount of money they are ready to spend to protect their personal data from disclosure).[15] Second, empirical studies point out that both the valuation of privacy and the valuation of personal data are extremely sensitive to contextual effects.

- A sixth way to value someone's personal data is to measure the price paid for an insurance to protect that data.[16]

Although an accurate measurement for data as an asset does not seem easily available, and although the variation in values amongst the various methodologies is significant, the OECD paper provided various inspiring approaches to tackle the issue and at least to estimate orders of magnitude.

Without attempting an empirical estimation of the value of data, Montes, Sand-Zantman and Valletti (2017) look into the value of personal information in online markets. They study how customer information and privacy affect the price-targeting behavior in online markets. Their theoretical model is based on

[12] For example, the implied market capitalisation per Facebook user has fluctuated between USD 40 and USD 300 at different times between 2006 and 2012.

[13] According to the OECD Report, for example, Facebook and Experian, two companies whose business models are based on personal data, have annual revenues per record/user of roughly USD 4-7 per year.

[14] An example is the security breach of the Sony's PlayStation Network and Sony Online Entertainment in 2011, which resulted in the exposure of 103 million records. According to Sony executives, this data breach will cost the company at least USD 171 million (USD 1.7 per record).

[15] This effect is not surprising. Behavioural economics often refers to situations like this as anchoring effect.

[16] Experian, a data broker, sells an identity-theft protection service called ProtectMyID for USD 155 per annum in the United States.

a range of specific hypotheses. They assume that data providers are the data brokers or data aggregators which act as intermediaries by collecting data from consumers and selling it to firms. This situation is not so common in the case of tech platforms such as Google, Uber or Facebook among many others. However, for the purpose of the current discussion, what is noteworthy to mention is that the three authors are amongst the first scholars attempting to give a meaning to the value of personal data, by means of a "privacy cost", defined as the cost consumers have to pay in order to "disappear" from a firm's database.

Finally, a discussion about the value of data cannot exclude the concept of "zero-price market" which has been developed by academic scholars such as John Newman or Daniel Rubinfeld. "Free" products became very popular along with widespread Internet adoption –but many of them are not truly free. Customers often trade their attention to advertisements or personal information to access zero-price products. The fact that in some cases the monetary price for providing a product or service is zero, does not mean that there is no value associated to that product or service. That value may be measurable in non-price parameters. Non-price competition is not new to antitrust laws. The most frequently discussed non-price elements of competition so far were quality, variety or innovation, but more and more often privacy becomes a relevant dimension.

Although the empirical measurement of the value of data has not been crystalized yet, it cannot be disputed at this stage that data is a valuable asset and an increasingly important parameter in the competitive assessment of firms' market power. In the following sections we will discuss some of the ways in which firms make use of data as an asset.

2. Network Effects and Barriers to Entry

If data is a valuable asset, and as such may provide a competitive advantage to the owners, the next question that may arise is whether the accumulation of data may give rise to market power and may create barriers to entry.

It has been argued that data-driven markets can become dominated by a few firms through network effects, which create barriers to entry. Economic theory distinguishes between direct and indirect network effects. Direct network effects arise when a consumer's utility from a product increases as others use the product. Telecommunications network are the classic example. Indirect network effects can be easily explained using the example of search engines: the more people use a search engine, the more trial-and-error experiments, the more likely the algorithms can learn of consumers preferences, the more relevant the

search results will likely be, which in turn attract others to use the search engine, and the positive feed-back continues (increasing returns to scale and scope).

The loop created by indirect network effects has the potential to reinforce incumbent's position but, at the same time, increase the product quality. In some circumstances, network effects may stimulate competition by allowing innovative entrants to grow rapidly. Competition authorities have looked into network effects with a cautious eye, investigating often if the presence of network effects does not lead to a tipping point beyond which dominance is the most likely outcome.

Network effects are currently pretty much central to the debate about whether online platforms are "unstoppable". Evans and Schmalensee (2018) attempt to debunk some myths related to the network effects, warning against slogans and advocating for evidence. They quote research showing a considerable churn in leadership for online platforms over periods shorter than a decade. This is largely due to the reverse network effects which are less mentioned in the debate. In the same way as networks can create exponential growth when additional customers attract more customers, networks can also lead to exponential decline, as each lost customer induces other customers to leave. The two authors explain that the apparent bias towards considering network effects potentially problematic comes from focusing on successful firms at a given point in time and concluding that they won it all and that they would not be displaced. A counterexample they provide is the case of Spotify that managed to become the leading source of digital music in the world, despite Apple having collected data through iTunes on more than 50 million users.

On a wider perspective, Rubinfeld and Gal (2016) explore entry barriers to big data markets and analyse some of their implication in the competitive analysis of such markets. They describe access barriers into all levels of the data value chain: collection, storage, analysis and usage and they also refer to all types of restraints: technological, legal or behavioural barriers.

As regards the collection of data, the two authors identify, among others, the following technological barriers: uniqueness of data or the gateways to it, economies of scale, scope and speed, network effects. The main legal barrier to entry in the collection of data are the data protection, privacy laws and the ownership rights and the main behavioural barrier is exclusivity, possibly leading to input foreclosure.

At the level of storage, there seem to be lower barriers to entry, especially due to technological advances allowing significant increases in storage space, such as, for example, the creation of cloud computing. Still, due to particular

structures and indexing inherent to the storage of a big amount of data, switching costs may arise when data is to be transferred to other systems or databases.

Regarding data analysis, Rubinfeld and Gal (2016) mention two important barriers to entry: data compatibility or interoperability and the analytical tools. The former refers to the intelligence used to rank the data and organise it in a certain relevant way. The latter is due the quality of algorithms used to process the big data. Certain firms own machine learning techniques that enable them to extract a unique value from the data, which is not replicable freely in the market.

Finally, even if data is accessible, either directly or through intermediaries, an additional barrier may limit the use of data when there are legal limitations designed to protect user's privacy.

After describing the four types of barriers to entry, Rubinfeld and Gal (2016) show how the characteristics of big data and entry barriers at each level of the value chain affect the competitive analysis. They admit that big data is non-rivalrous and collecting it may not prevent others from collecting identical data. However, this observation should not lead automatically to claims of low barriers to entry, precisely because data collection is only one of the stages in the data value chain. And entry barriers can create competitive effects, such as exclusionary conduct, similar to those in traditional markets. However, the two authors show that big data markets, due to some unique characteristics, may exhibit twists on the regular analysis, which in turn, may affect theories of harm. Among others, they mention the following characteristics:

- Data is multi-dimensional, any of the four or five 'Vs" could enhance or lower barriers to entry, hence a market specific approach is needed to determine the direction of the effect.
- The non-rivalrous nature of data does not necessarily mean that the collection, organisation, storage or analysis cannot transform it into a private good. However, if the barriers to entry are structural and sharing the data is socially beneficial, a regulatory solution may be appropriate, possibly by requirements that the data is made available at FRAND terms (fair, reasonable and non-discriminatory), such as is the case often with patents.
- When data is an input, the analysis of related markets is also necessary. Especially in the case of free on-line services, both the market for the collection of data and the advertising market are affected.

- Big data might strengthen price discrimination when it contains information regarding consumer preferences.[17]

Another paper assessing barriers to entry in big data markets is Lambrecht and Tucker (2017). They took a different perspective and toned down considerably the value of data and its potential to constitute a barrier to entry. They argue that, for a firm resource (including the data) to be a source of competitive advantage, the resource has to be inimitable, rare, valuable and non-substitutable. Their analysis suggests that data is not inimitable or rare, it has substitutes and it is not valuable by itself.

Data is considered inimitable when no firm is be able to replicate it. In the authors' view, big data does not possess the features to make it inimitable. Data is non-rivalrous (*i.e.,* its consumption does not decrease its availability to others) and it has near-zero marginal cost of production and distribution making it possible to be resold. Commercially available data has a broad coverage. Furthermore, they claim that data is not rare and tools for gathering big data are becoming more and more common. Many consumers use multi-homing (a single consumer commonly uses different digital services), similar pieces of information are therefore available to different companies.

The two authors also claim that the characteristics of big data potentially undermine their value as a competitive advantages. Data is often unstructured and establishing causal relationships is difficult within large pools of overlapping observations. Firms need to move from observational correlations to the identification of the relationships that should guide their decision making.[18] Moving from correlation to causal relationships requires the design of either experiments, which often do not require such huge amount of data, or algorithms that are better at dealing with the data. They provide an interesting reference showing that often it is not the size of that data that matters but the machine-learning algorithm used to determine the quality of the results. To support this claim, they quote Pilaszy and Tikk (2009)[19] who show that ten movie ratings alone are more helpful than extensive metadata to predict preferences for movies.

Finally, Lambrecht and Tucker (2017) claim that data is substitutable. History shows that in different data-intensive industries, new entrants have

[17] This issues will be expanded later in this chapter.

[18] Note that Marr (2016) has claimed otherwise.

[19] Available at: https://www.researchgate.net/publication/221141047_Recommending_new_movies_Even_a_few_ratings_are_more_valuable_than_metadata

been successful even in the presence of incumbents owning big data thanks to a superior ability to understand and meet customer needs. Examples of successful entry in different markets are numerous: WhatsApp and Snapchat (communications industry); Uber and AirBnB (sharing economy); Tinder, where personalized experience is key. They infer that, in order to gain a sustainable competitive advantage from big data, firms are required to develop the right managerial, engineering and analytical skills to transform data into valuable and actionable knowledge.

This conclusion is very much in line with Hal Varian's (Google's Chief Economist) view according to which there are decreasing returns to scale in data, meaning that each additional piece of data is somehow less valuable and at some point, collecting more does not add anything. What matters more, he says, is the quality of the algorithms that crunch the data and the talent a firm has hired to develop them. Google's success " is about recipes, not ingredients". Varian's view contrasts with the opinion of Glen Weyl, Microsoft researcher, who believes that algorithms are self-teaching –the more and the fresher data they are fed, the better.[20]

Whereas big data is unarguably an important asset for firms in the competition game, it is also clear that data in itself is not sufficient to place businesses above their competitors. Data has to be collected by powerful machines, processed by intelligent algorithms and used by ingenious minds in order to provide a competitive advantage. Furthermore, establishing whether this competitive advantage is used in a potential anti-competitive way requires a good understanding of the specific big data market and its characteristics. The literature cited in this section provides good guidance on the elements to be assessed for this purpose.

3. Data as an Input: Is it an Essential Facility?

In 2016, the EU Competition Commissioner, Margrethe Vestager addressed the issue of data uniqueness, which may make it an essential input:[21] "The problem for competition isn't just that one company holds a lot of data. The problem comes if that data really is unique, and can't be duplicated by anyone else. But really unique data might not be that common.

[20] See *The Economist,* 6th May 2017.

[21] See: https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/making-data-work-us_en

That doesn't mean the amount or type of data that a company controls can never create a problem.[22] […] we shouldn't be suspicious of every company which holds a valuable set of data. But we do need to keep a close eye on whether companies control unique data, which no one else can get hold of, and can use it to shut their rivals out of the market."

Jurisprudence has defined the conditions for an input or a resource, in general, to be considered an essential facility. Colangelo and Maggiolino (2017) review that jurisprudence and state that "a facility is essential when it –and only it– serves to offer a specific product or service; that is, when there is a cause-effect relationship between the facility and the good or the service that the rival wants to realize by using that facility". In the case of data, the two authors question that some specific data can be considered an essential facility because it is not even clear what information could be obtained from it. The authors conclude that, if this doctrine needs to be applied in this context, it should refer to the information and not to the data itself. It would be more accurate to focus on the data processing and information production steps of the data value chain instead of the process of data accumulation.[23]

M. Cole (2018) suggests that there are strong parallels between the approach to assessing the potential for data to foreclose access, and the approach to assessing the potential for the intellectual property to foreclose. In such a setting, the set of laws to govern the ownership and exchange of data may be similar to licensing of patents.

It has been already advocated by several scholars that data sharing can be an important tool in achieving efficiency and avoiding market dominance. Using a theoretical model, Prüfer and Schottmüller (2017) show that in a market with at least two firms with market power, and where innovation investments are decided repeatedly over time, the market will tip under very mild conditions, meaning it will go towards monopoly. Interestingly, they show that market tipping can be avoided if competitors share their user information. Moreover, they also show that a dominant firm's incentives to innovate further do not decline after forced sharing of user information.

This finding can have potential important policy implications, as it suggests that data sharing can help reducing the barriers to entry into the market

[22] Vestager's speech provides an interesting antitrust case where data was considered unique: "In 2014, the French competition authority ordered GDF Suez, a French energy supplier, to share a rather traditional type of data –part of its customer list– with its rivals. That list was special because it related to regulated tariffs, which only GDF Suez could legally offer. And the French competition authority was concerned that GDF Suez might have misused that list, which it had because of its monopoly, to sell energy in the part of the market that was open to competition."

[23] Their view in this respect is in line with Rubinfeld and Gal (2016).

and thus be a powerful tool in avoiding the market power tipping towards dominance.[24]

In circumstances where data is considered unique, and treated like an essential facility similarly to an essential patent, competition authorities and regulators should trade-off the benefits to competition and consumers from data sharing with the protection of the tech companies' business models, which rely heavily on data and information. Similar trade-offs made the object of recent mergers where the parties advocated for concentration increasing the incentives to innovate whereas competition agencies viewed competition as the main trigger to innovation.[25] Data intermediaries and data brokers may play a certain role in distribution of data in a commercially viable way.

An interesting example of data sharing is Google Trends, where Google reveals data about people's searches. Data is aggregated enough such that people cannot be identified, yet detailed enough to enable interesting analyses. In his book, Seth Stephens-Davidowitz (2017) provides a comprehensive set of interesting stories and anecdotes purely based on the data made publicly accessible by Google.

Finally, an adjacent issue to data sharing is data portability. Article 20 of the GDPR[26] creates a new right to data portability, which is closely related to the right of access but differs from it in many ways. It allows for data subjects to receive the personal data that they have provided to a controller, in a structured, commonly used and machine-readable format, and to transmit those data to another data controller. The purpose of this new right is to empower the data subject and give him/her more control over the personal data concerning him or her. Since it allows the direct transmission of personal data from one data controller to another, the right to data portability is also an important tool that will support the free flow of personal data in the EU and foster competition between controllers. It will facilitate switching between different service providers, and will therefore foster the development of new services in the context of the digital single market strategy.

---

[24] Rubinfeld and Gal have also suggested data sharing as a possible regulatory intervention and they even go further and propose a FRAND-type agreement, as in the patent licensing theory.

[25] Similar trade-offs made the object of recent mergers where the parties advocated for concentration increasing the incentives to innovate whereas competition agencies viewed competition as the main trigger to innovation. See for example Shapiro(2016) for a detailed analysis of the classic Schumpeter-Arrow debate.

[26] Reference General Data Protection Regulation, http://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf

IV. COLLUSION

One competition policy area where the development of the digital economy has spurred increasing debate is collusion. Most debate so far is hypothetical, based on the theory of collusion and looking at whether the algorithmic competition may have an impact on that factors that facilitate collusion.

Ezrachi and Stuke (2016) note that Big Data and Big Analytics –by increasing the speed of communicating price changes, detecting any cheating or deviations, and punishing such deviations– can provide new and enhanced means to foster collusion. They consider four scenarios in which computer algorithms may promote collusion:

- The first scenario –Messenger– concerns humans' agreeing to collude and using computers to execute their will.[27]

- The second scenario –Hub and Spoke– considers the use of a single pricing algorithm to determine the market price charged by numerous users; in this framework, a cluster of similar vertical agreements with many of the industry's competitors may give rise to a classic hub-and-spoke conspiracy, whereby the algorithm developer, as the hub, helps orchestrate industry-wide collusion, leading to higher prices (to illustrate this scenario, the authors use the Uber technology as an example)

- The third scenario –the Predictable Agent– explores how we are shifting from a world where executives expressly collude in smoked-filled hotel rooms, to a world where pricing algorithms act as predictable agents and continuously monitor and adjust to each other's prices and market data. The result is algorithm enhanced conscious parallelism. However, this is a form of tacit collusion which comes along with its enforcement challenges.

- Finally, the most challenging scenario in the authors' opinion –the Digital Eye– represents a situation where the computers, in learning by doing, determine independently the means to maximize profits. Artificial intelligence operating in enhanced market transparency leads to an anticompetitive outcome, with no evidence of any anti-competitive agreement or intent. In this case, the authors acknowledge that not only the harm but also the illegality are very difficult to prove.

Economic theory has identified certain industry characteristics that have an impact on the likelihood of reaching and sustaining collusion, such as: the

[27] Various case references are provided in their book.

number of firms, barriers to entry, transparency, frequency of interactions, asymmetries, innovation, among others. According to OECD (2017), the increasing use of algorithms enhance some of these characteristics, making collusion more likely, as described below.

While a high number of firms was an indicator of a difficult environment to collude in traditional industries, the use of algorithms could allow coordination and monitoring of a larger number of firms. As regards barriers to entry, the impact of algorithms is ambiguous. On the one hand, algorithms can be used to identify any market threats very fast, allowing incumbents to pre-emptively acquire any potential competitors or to react aggressively to market entry. On the other hand, the increasing availability of online data resulting from the use of algorithms may provide useful market information to potential entrants and improve certainty, which could reduce entry costs.

Furthermore, algorithms are very likely to increase both transparency in the market as well as the frequency of interactions, which make industries more prone to collusion. The increase of market transparency is not only a result of more data being available, but also of the ability of algorithms to make predictions and to reduce strategic uncertainty. Complex algorithms with powerful data mining capacity are in a better place to distinguish between intentional deviations from collusion and natural reactions to changes in market conditions.

With respect to the frequency of interaction, the digital economy has revolutionised the speed at which firms can make business decisions. Prices may be updated in real-time, allowing for an immediate retaliation to deviations from collusion.

Interestingly, the OECD (2017) paper describes how some supply characteristics of digital markets may counterbalance the enhanced risk of collusion resulting from more transparent markets. One of the most relevant supply-side characteristics is innovation. Algorithms are naturally an important source of innovation, allowing companies to develop new business models and extract more information from data, in order to respond to customers' needs. In industries where the algorithm is a source of competitive advantage, companies may face a greater competitive pressure to develop the best-performing algorithm. Similarly, if algorithms allow companies to differentiate their services or the production process in such a way that leads to asymmetries on the supply side, collusion might be again harder to sustain, due to the inherent difficulties of finding a focal point to coordinate and as a result of the low incentives for the low-cost firms to collude.

Finally, the OECD (2017) paper discusses policy implications of these potential challenges that algorithms pose on the standards analysis of collusion. For this purpose, OECD refers to the standard distinction between tacit and explicit collusion. If algorithms amplify conduct which is already covered under the current legal framework (*i.e.,* explicit collusion) the discussion is rather straightforward, as algorithms ought to be assessed together with the main infringement that they help enforcing. While detecting the existence of an infringement and proving such an infringement might still be complex because of the presence of an algorithm, agencies can nevertheless rely on existing rules on anti-competitive agreements, concerted practices and facilitating practices, which offer agencies a framework to assess algorithms either on their own or as practices ancillary to a main infringement.

However, algorithms may, to some extent, create new risks related to behaviours not covered by the current antitrust rules. This is the issue of algorithms achieving a tacitly collusive equilibrium without any need for contact between competitors or without putting in place any facilitating practice. Faced with this challenge, OECD poses the question of whether the notion of agreement should be revisited. Most probably, a more clearer definition of agreement could not only reduce uncertainty by helping businesses understanding which practices are illegal and which ones are acceptable, but also to potentially address some of the concerns related to algorithmic collusion. For the moment, OECD suggests that some competition enforcers may have the possibility to rely on legal standards such as "unfair competition" which provide more flexibility.

In addition, OECD suggests possible alternative approaches to assess algorithmic collusion: market studies and market investigations to inform possible regulatory interventions or ex ante merger control. Finally, since the since algorithms can result in multiple other market failures (*i.e.:* information asymmetries resulting from lack of algorithmic transparency, data-driven barriers to entry, spillovers associated with information and knowledge), an increasing attention has been given to the potential need for a regulatory reform in the digital economy. The OECD paper mentions a few regulatory approaches that might be considered in the future to tackle algorithmic collusion, such as price regulation, policies to make tacit collusion unstable and rules on algorithm design. However, given the multi-dimensional nature of algorithms, policy approaches should be developed in co-operation with competition law enforcers, consumer protection authorities, data protection agencies, relevant sectorial regulators and organisations of computer science with expertise in deep learning. Both lack of intervention and overregulation could pose serious costs on society, hence further actions should be subject to in-depth assessment.

A final issue that has been raised in relation to the algorithmic collusion is the anti-trust liability. EU Competition Commissioner has stated very clearly

that "companies can't escape responsibility for collusion by hiding behind a computer program."[28] Furthermore, she advised that "compliance with the rules –the competition rules, for instance– should be built into those algorithms by design. So that even if we don't know exactly how they make their decisions, we can be confident that algorithms will act like good citizens."[29]

V. ABUSE OF DOMINANT POSITION

This section will discuss some general considerations regarding specific potential abuses related to big data and two cases. Big tech companies are in the radar of competition agencies, but theories of harm based strictly on big data are still at an incipient phase. We will briefly describe two cases that have certain relevance in the big data discussion: Google and Facebook.

1. Behavioral Discrimination and Personalized Pricing

One potential unilateral theory of harm that has been discussed since firms have been increasingly using algorithms in their pricing decisions is the behavioral discrimination and personalized pricing.

A producer price-discriminates when two units of the same good are sold at different prices, either to the same consumer or to different consumers. For price discrimination to be possible, two main ingredients are necessary: firms must be able to sort consumers[30] and arbitrage should be absent.

Article 102 (c) TFEU[31] considers as an abusive behaviour the fact that a dominant firm applies: "dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage". However, Commission's 102 Guidance[32] does not provide any further direction on how to assess cases of price discrimination.

---

[28] https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/bundeskartellamt-18th-conference-competition-berlin-16-march-2017_en

[29] https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/helping-people-cope-technological-change_en

[30] Depending on the information available to firms that is used to sort consumers, price discrimination can be: first-degree (perfect) when the valuation of every consumer is known and the firm charges everyone a different price (the maximum everyone would pay); second-degree when firms offer different deals and consumers "self-select" themselves: and third-degree price discrimination when firms charge different prices to consumers having different (observable) characteristics.

[31] Treaty for the Functioning of the European Union.

[32] Communication from the Commission: Guidance on its enforcement priorities in applying Article 82 of the EC Treaty to abusive exclusionary conduct by dominant undertakings. OJ C 45, 24.2.2009, p. 7–20.

Economic theory has shown that welfare effects of price discrimination are ambiguous. Discrimination allows firms to reduce prices to categories of consumers who would not buy otherwise. In principle if output does not increase, price discrimination reduces consumer welfare. However, if output increases, consumer welfare may increase with discrimination.

Ezrachi and Stucke (2016) put forward a possible theory of harm around price discrimination enhanced by the use of algorithmic pricing (see also Rubinfeld and Gal, 2016). They explore how personalization of our online environment through search inquiries, past purchase or e-mails affects the dynamics of competition and consumer welfare. They acknowledge how behavioral advertising, personalized product offerings and targeted pricing can help reduce consumers' search costs and save their time. However, the authors note that behavioral discrimination can reduce their welfare also as "individualization" does not stop at promotions but it affects pricing decisions too, so that the more vulnerable end up paying more. According to Ezrachi and Stucke (2016), personalized pricing is possible due to informational asymmetries between the discriminating firm and its customers and also between the firm and its competitors.

In a Report commissioned by the Centre on Regulation in Europe, Bourreau *et al.* (2017), acknowledging the ambiguous effects of price discrimination on consumer welfare, state that there is no rationale for banning personalized pricing per se. However, one concern could be that price discrimination is used as a monopolization device, for example if an incumbent firm pre-empts entry in a given market or consumer segment by setting very low prices or loyalty discount in this market. This type of concern could be aggravated if possibilities of price discrimination hinge on detailed consumer data, and incumbent firms have exclusive access to this consumer data. Their main policy recommendation would be that personalized pricing strategies, if they exist, should be transparent to ensure consumers' trust in online markets. This also requires an effective application of consumer protection, for instance by monitoring online prices by consumer protection agencies upon complaints.

2. Exclusionary Conduct: Google Shopping[33]

According to the European Commission's press release, Google has abused its market dominance as a search engine by giving an illegal advantage to another Google product, its comparison shopping service. The fine for breaching EU antitrust rules was €2.42 billion. This case spurred numerous discussions in the

[33] See Commission Decision of 27.6.2017 in the case AT.39740 - Google Search (Shopping).

competition community. We will not comment on this general debate here but will rather focus on the elements from the Commission decision that may offer some guidance on the role of data on providing market power.

Various Commission documents describe the facts of the case.[34] Google's flagship product is the Google search engine, which provides search results to consumers. In 2004 Google entered the separate market of comparison shopping in Europe. Comparison shopping services rely to a large extent on traffic to be competitive. Google's search engine is an important source of traffic for comparison shopping services.

Commission's decision concluded that Google is dominant in each national market for general internet search throughout the European Economic Area, because Google has very high market shares and there are also high barriers to entry, in part because of network effects. Furthermore, the Commission asserted that Google has abused its market dominance in general internet search by giving a separate Google product an illegal advantage in the separate comparison shopping market by two means:

- Google has systematically given prominent placement to its own comparison shopping service.

- Google has demoted rival comparison shopping services in its search results.

The Commission found that Google's conduct has potential anti-competitive effects. Firstly, it could foreclose competing comparison shopping services, which may lead to higher fees for merchants, higher prices for consumers, and less innovation. Secondly, Google's conduct is likely to reduce the ability of consumers to access the most relevant comparison shopping services.

The main part of the decision where the Commission focuses on data is related to the barriers to entry that may support Google's dominance. For the scope of this chapter's discussion, we will pinpoint to the barriers to entry the Commission found in its assessment.

Paragraph 287 of the Commission Decision states that "because a general search service uses search data to refine the relevance of its general search results pages, it needs to receive a certain volume of queries in order to compete viably. The greater the number of queries a general search service receives, the

[34] See: http://ec.europa.eu/competition/elojade/isef/case_details.cfm?proc_code=1_39740

quicker it is able to detect a change in user behaviour patterns and update and improve its relevance."

Although the Commission recognizes in a subsequent paragraph (289) that there may be diminishing returns to scale in terms of improvements in relevance once the volume of queries a general search service receives exceeds a certain volume, the relevance of scale is not called into question as a general search service has to receive at least a certain minimum volume of queries in order to compete viably. Hence the Commission considered that data itself can be considered a barrier to entry, especially taking into account the 'volume' characteristics.

Further paragraphs refer to the barriers to entry created by the positive feedback effects on both sides of the two-sided platform formed by general search services and online search advertising. As regards the online search advertising, the higher the number of users of a general search service, the greater the likelihood that a given search advertisement is matched to a user and converted into a sale. This in turn increases the price that a general search engine can charge advertisers if their search advertisements are clicked on.[35]

As regards the positive feedback effects on the general search side of the platform, the Commission believes they are derived from both direct and indirect network effects. The direct effects stem from the fact that a substantial minority of users of a general search service derive a benefit from such advertisements.[36] The indirect network effects stem from the link between the attractiveness of the online search advertising side of the platform and the revenue of the platform. The higher the number of advertisers using an online search advertising service, the higher the revenue of the general search engine platform; revenue which can be reinvested in the maintenance and improvement of the general search service so as to attract more users.[37]

The discussions around network effects and barriers to entry in the Commission decision were used to establish dominance. Google Shopping is not a case based primarily on a big data theory of harm. Fumagalli, Motta and Galgano (2018), even if writing before the decision was public, frame clearly possible theories of harm, confirmed later by the decision. They explain how Google, to the extent that it is dominant in the upstream market of search services (input market), could engage in foreclosure strategies by denying access to such an input to its comparison shopping rivals. However the input

[35] See paragraph 293 of the Google Decision.

[36] See paragraph 295 of the Google Decision.

[37] See paragraph 296 of the Google Decision.

was not the data in this case. A competition case based primarily on a big data exclusionary theory of harm has not yet been pursued.

3. Exploitative Conduct: Facebook[38]

Whereas the Google case was run by the European Commission from an exclusionary conduct angle, another interesting investigation, this time by the German Competition Authority, concerns Facebook and is tackling a possible exploitative abuse. The case is still ongoing at the time of drafting this chapter, but in December 2017 the German Competition Authority has sent Facebook its preliminary assessment. The information contained in this section is based on press releases.

The authority assumes that Facebook is dominant on the German market for social networks and holds the view that Facebook is abusing this dominant position by making the use of its social network conditional on its being allowed to limitlessly amass every kind of data generated by using third-party websites and merge it with the user's Facebook account. These third-party sites include firstly services owned by Facebook such as WhatsApp or Instagram, and secondly websites and apps of other operators with embedded Facebook application.

The authority is concerned that users cannot switch to other social networks and that participation in Facebook's network is conditional on registration and unrestricted approval of its terms of service. According to authority's preliminary assessment, Facebook's terms of service violate data protection provisions to the disadvantage of its users as it cannot be assumed that users effectively consent to this form of data collection and processing. The focus of the investigation is on the collection of data from third party websites and not on the social network itself. The German authority states that users cannot expect data which is generated when they use services other than Facebook to be added to their Facebook account to this extent and is closely cooperating with data protection authorities as regards the data protection aspects of the case.

This is the first time a competition authority initiated a potential abuse of dominance on the ground of an infringement of data protection rules. According to Andreas Mundt, the president of the German authority, "when data is called the new currency of the digital age, then the relationship to competition law is obvious."[39] The German competition authority is assessing Facebook's alleged

[38] See press release from the German Competition Authority at: https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2017/19_12_2017_Facebook.html

[39] https://www.financialexpress.com/industry/technology/facebooks-new-nemesis-is-a-besuited-german-antitrust-watchdog-named-andreas-mundt/982665/

abuse of dominance in the market for social networks through the lens of excessive pricing. Officials from the German Competition Authority[40] talk about a possible "excessive data processing" case, as Facebook users have to accept a lot of data processing as a precondition to use Facebook, leading to them paying with their data.

The German official said that exploitative business terms can constitute an abuse under German law and one benchmark for such an exploitation can be a breach of data protection principles, when there is no freely given consent. The German authority is aware though that a social network needs efficient data-based product design to prosper and that users probably expect a certain processing of their data as payment to use a free service.

This case is challenging in many respects. First, defining a data market in order to determine dominance requires a thorough understanding of an unprecedented product that has multiple dimensions. Graef (2015) states that under current competition law standards, a correct market definition requires the existence of supply and demand for the product or service. She follows that, the relevant market for online services such as search engines, social networks and e-commerce platforms cannot take data as object as long as there is no economic transaction between the respective providers and users for data, and the providers of these online platforms do not sell or trade data to third parties.

Second, establishing an exploitative abuse is already challenging when the main parameter is the price, hence finding a benchmark for "excessive data" needs very careful, legally sound and thorough arguments that should not undermine the business model of platforms such as Facebook that grow and deliver value to consumers precisely by using big data.

Exploitative abuses are not that common and they are difficult to argue, mainly due to the difficulty in finding a reasonable benchmark. The Commission Guidelines on the application of Article 102 of the TFEU[41] does not elaborate on exploitative abuses, although the Article 102 itself stipulates these practices mentioned as "unfair trading conditions." Previous cases of exploitative abuses concerned excessive pricing, often in the context of the pharmaceutical companies. Excessive pricing has not been applied so far in the context of an antitrust abuse in the digital economy.

Finally, tackling abuses of privacy under competition law is also unprecedented. In section six we will briefly discuss a few merger cases where

[40] See PaRR article quoting Krueger: https://app.parr-global.com/intelligence/view/prime-2604376

[41] Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52009XC0224(01)&from=EN

the European Commission touched the privacy issue as a possible non-price parameter of competition.

VI. MERGERS

Although the recent wave of data-driven mergers shows how much companies value data, the European Commission has not intervened significantly in transactions that involved big data. However, interesting insights can be derived from its analysis in these cases.

The Economist compiled a list of selected data-driven deals in the past five years and the numbers are impressive. The two biggest in value are Facebook/WhatsApp and Microsoft/LinkedIn. We will briefly describe European Commission's assessment of these two mergers.

TABLE 1

DATA-DRIVEN DEALS, SELECTED

Company	*Target company (Date)*	*Value of deal, $bn*	*Business*
	Instagram (2012)	1.0	Photo sharing
	WhatsApp (2014)	22.0	Ext/photo messaging
Alphabet	Waze (2013)	1.2	Mapping and navigation
IBM	The Weather Company (2015)	2.0	Meteorology
	Truven Health Analytics (2016)	2.6	Health care
intel	Mobileye (2017)	15.3	Self-driving cars
Microsoft	SwiftKey (2016)	0.25	Keyboard/artificial intelligence
	LinkedIn (2016)	26.2	Business networking
ORACLE	BlueKai (2014)	0.4	Cloud data platform
	Datalogix (2014)	1.0	Marketing

Source: The Economist, 6 May 2017, Briefing – The Data Economy.

In 2017 companies spent around $22bn on artificial intelligence related mergers and acquisitions, about 26 times more than in 2015.[42]

Despite the significant value of these transactions, some of them involve companies that do not have important revenues because many of them are

[42] *The Economist,* 31st March 2018.

very young. However, they possess other assets such as data or ideas, which are extremely valuable for other big tech companies. For this reasons, it may happen that the current thresholds for merger notification based mainly on revenues are not met and consequently competition authorities do not investigate those transactions. This was the case of the very recent proposed acquisition by Apple of Shazam, which did not meet the threshold at EU level. However, that transaction ended up with the European Commission following a request by various competition authorities, as we will discuss at the end of this section.[43]

Faced with this issue, various national competition authorities in Europe, such as Germany and Austria,[44] have introduced the value of the transaction amongst the criteria to notify a merger, enabling them to scrutinize also acquisitions of relatively small companies valued highly by the market. Including the value of the transaction instead of lowering the revenue threshold might filter relevant mergers without increasing the burden of the competition authorities to assess many more transactions of smaller companies.

Acquisition of small valuable companies raises a more fundamental question: should acquiring start-ups be a legitimate growth strategy? Alternatively, on the flip side, should exit via being acquired be a legitimate entrepreneurial strategy for start-ups? Ideas, both in the form of innovation or in the form of data collection can be acquired either through take-overs or through internal developments. Acquisitions provide a ready-made product whereas innovation has a high probability of failure, which, of course, has a price.

Competition enforcers rightly fear that acquisition of ideas has the potential anti-competitive effect of foreclosing the market, which may lead to harm to consumers. However, synergies in this type of transactions may bring also value to consumers. This trade-off between concentration and competition is not new and has been debated in various mergers focusing on innovation. The discussion is based on the classic Schumpeter versus Arrow debate (see, for example Shapiro, 2016), the former scholar advocating for market power to increase innovation whereas the latter for competition to spur incentives to become better than rivals. In such a space, it is possible to draw parallels between the acquisition of innovation and acquisition of data. However, a forward-looking assessment in such an uncertain environment is not an easy task.

We end this section by briefly describing two mergers investigated by the European Commission where data issues played a certain role. Even though

[43] See European Commission press release available at: http://europa.eu/rapid/press-release_IP-18-664_en.htm

[44] Other European countries expressed intention to follow.

all companies involved in these transactions were high tech companies heavily relying on big data, the Commission did not find significant concerns related to the data. At the time of drafting this chapter, the Commission has announced the opening of the in-depth investigation in the acquisition of Shazam by Apple, two significant and well known players in the digital music industry that are mainly active in complementary business areas, namely music streaming service (Apple Music, number two in Europe) and music recognition app for mobile devices where Shazam is the market leader. At this stage, the Commission is concerned that, following the takeover of Shazam, Apple would obtain access to commercially sensitive data about customers of its competitors for the provision of music streaming services in the EEA, which could allow Apple to directly target its competitors' customers and encourage them to switch to Apple Music. As a result, competing music streaming services could be put at a competitive disadvantage. In addition, while at this stage the Commission does not consider Shazam as a key entry point for music streaming services, it will also further investigate whether Apple Music's competitors would be harmed if Apple, after the transaction, were to discontinue referrals from the Shazam app to them.[45] It remains to be seen whether in this case the Commission will take a step further in the assessment of data-driven markets.

1. Facebook/WhatsApp[46]

In 2014, the European Commission approved the acquisition of WhatsApp by Facebook. Both companies offer applications for smartphones which allow consumers to communicate by sending text, photo, voice and video messages. The Commission found that Facebook Messenger and WhatsApp are not close competitors and that consumers would continue to have a wide choice of alternative consumer communications apps after the transaction.

The Commission's investigation focused on three areas: (i) consumer communications services, (ii) social networking services, and (iii) online advertising services.

As regards consumer communications services, the Commission focussed its assessment on apps for smartphones, as WhatsApp is not available for other devices. The Commission found that Facebook Messenger and WhatsApp are not close competitors. For WhatsApp, access to the service is provided through

[45] See European Commission's press release, available at: http://europa.eu/rapid/press-release_IP-18-3505_en.htm

[46] See European Commission Decision in case M. 7217, available at http://ec.europa.eu/competition/mergers/cases/decisions/m7217_20141003_20310_3962132_EN.pdf

phone numbers while for Facebook Messenger, a Facebook profile is required. Furthermore, this is a very dynamic market with several competing apps available on the market, such as Line, Viber, iMessage, Telegram, WeChat and Google Hangouts. Although the Commission found that consumer communications apps market is characterised by network effects, a number of factors were recognised to mitigate the network effects in this particular case. Indeed, the Commission found that the consumer communications apps market is fast growing and characterised by short innovation cycles in which market positions are often reshuffled. Moreover, launching a new app is fairly easy and does not require significant time and investment. Finally, customers can and do use multiple apps at the same time and can easily switch from one to another.

As regards social networking services, the Commission also found that the parties are, if anything, distant competitors.

Finally, although WhatsApp is not active in online advertising, the Commission examined whether the transaction could strengthen Facebook's position in that market and hamper competition. In particular, the Commission examined the possibility that Facebook could (i) introduce advertising on WhatsApp, and/or (ii) use WhatsApp as a potential source of user data for improving the targeting of Facebook's advertisements. The Commission concluded that, a large amount of internet user data that are valuable for advertising purposes are not within Facebook's exclusive control, hence the merger will not impact negatively the market for advertisers.

In the context of this investigation, the Commission analysed potential data concentration issues only to the extent that it could hamper competition in the online advertising market. Paragraph 164 of the Decision states: "Any privacy-related concerns flowing from the increased concentration of data within the control of Facebook as a result of the Transaction do not fall within the scope of the EU competition law rules but within the scope of the EU data protection rules."

Several commentators, including Stucke and Grunes (2016) among others, stress nevertheless that privacy is an important factor of non-price competition. We have already seen that the German Competition Authority's approach in the Facebook abuse case is also inclined towards defining privacy abuses of dominance within the realm of competition law.

The two authors state that the Commission erred in considering that the concerns of one firm controlling so much data were strictly a privacy issue, not a competition issue. They explain the difference between the two business models of Facebook and Whatsapp from the perspective the price/privacy trade-

offs. Whatsapp charged users a nominal fee for the service and promised not to collect data, whereas Facebook provides the service for free but harvests consumers data instead, charging advertisers for supporting them in targeting their adds.

The Commission cleared the merger on the assumption that Facebook would be unable to establish reliable automated matching between Facebook users' accounts and WhatsApp users' accounts. However, a few years later, the Commission found the contrary and fined Facebook for providing incorrect or misleading information during the merger investigation.[47]

2. Microsoft/LinkedIn[48]

In 2016 the European Commission has approved the acquisition of LinkedIn by Microsoft subject to certain commitments aimed at preserving competition between professional social networks in Europe. Microsoft and LinkedIn are mainly active in complementary business areas, except for minor overlaps in online advertising.

Microsoft develops, licenses, and supports software products, services and devices. Microsoft also provides other software solutions, including customer relationship management (branded "Dynamics"), which is a type of software used by businesses to manage their sales, marketing and customer support activities.

LinkedIn operates the internet-based social networking service that focuses on promoting professional connections. Professional social network services are offered as free of charge, basic subscriptions or premium subscriptions. Among premium subscriptions, LinkedIn offers a sales intelligence solution for businesses branded "Sales Navigator." This product grants access to a subset of the entire LinkedIn database that can be purchased by businesses that also buy customer relationship management solutions.

The Commission assessed both horizontal and non-horizontal effects stemming from the transaction. Potential horizontal effects could arise in relation to the online advertising services. However, given their very limited combined market share in the EEA, as well as the fragmented nature of the market, the Commission excluded any competition concerns in this area.

[47] See European Commission Decision in case M.8228, available at: http://ec.europa.eu/competition/mergers/cases/decisions/m8228_493_3.pdf

[48] See European Commission Decision in case M.8124, available at: http://ec.europa.eu/competition/mergers/cases/decisions/m8124_1349_5.pdf

Related to this horizontal theory of harm, the Commission also looked at the concentration of the parties' user data (essentially consisting of personal information, such as information about an individual's job, career history and professional connections, and/or her or his email or other contacts, search behaviour etc.) that can be used for advertising purposes. No concerns were found in this respect either, for the following reasons: (i) the combination of data was subject to Data Protection rules; (ii) Microsoft and LinkedIn do not make available in general their data to third parties for advertising purposes; (iii) the combination of their respective datasets does not appear to result in raising the barriers to entry/expansion for other players in this space, as there will continue to be a large amount of internet user data that are valuable for advertising purposes and that are not within Microsoft's exclusive control and (iv) the Parties are small market players and compete with each other only to a very limited extent in online advertising. Competition Commissioner, M. Vestager stated in relation to the approval of this merger: "We had to look closely at exactly what data was involved, and how it would really affect competition. After all, controlling a lot of data isn't such a big issue, if others can easily get hold of the same information, from their own customers or simply by buying it in the market. And that's just what we found in the case of Microsoft and LinkedIn –that even after the merger, other companies would still have access to comparable or even better data than LinkedIn."

The non-horizontal theory of harm in this case focused on the professional social network services market and the customer relationship management software solutions market. As regards the latter, the Commission did not find significant concerns.

In the case of the professional social network services, the Commission was concerned that Microsoft would pre-install LinkedIn on all Windows PCs, integrate LinkedIn into Microsoft Office and combine LinkedIn's and Microsoft's user databases. Furthermore, to the extent that these foreclosure effects would lead to the marginalisation of an existing competitor which offers a greater degree of privacy protection to users than LinkedIn (or make the entry of any such competitor more difficult), the Commission considered that the transaction would also restrict consumer choice in relation to this important parameter of competition when choosing a professional social network. Even though the Commission's view in this case was that privacy related concerns as such do not fall within the scope of EU competition law, it considered that they can be taken into account in the competition assessment to the extent that consumers see it as a significant factor of quality, and the merging parties compete with each other on this factor.[49] In fact, Commission's investigation revealed that,

[49] This is consistent with Commission's approach in Facebook/WhatsApp merger.

today, in Germany and Austria, Xing seems to offer a greater degree of privacy protection than LinkedIn.[50]

Microsoft offered the following commitments that addressed the competition concerns identified by the Commission in this area.[51]

- Ensure that PC manufacturers and distributors would be free not to install LinkedIn on Windows and allow users to remove LinkedIn from Windows should PC manufacturers and distributors decide to preinstall it;

- allow competing professional social network service providers to maintain current levels of interoperability with Microsoft's Office suite of products through the so-called Office add-in program and Office application programming interfaces;

- grant competing professional social network service providers access to "Microsoft Graph", a gateway for software developers. It is used to build applications and services that can, subject to user consent, access data stored in the Microsoft cloud, such as contact information, calendar information, emails, etc. Software developers can potentially use this data to drive subscribers and usage to their professional social networks.

VII. CONCLUSION

This chapter has reviewed current competition policy issues in the world of big data. A survey of economic literature, policy papers and competition cases revealed divergent views as regards the role and effects of big data and machine learning on consumers and society in general.

We discuss some of the relevant questions of this debate. We believe that, at this stage, there is no controversy on whether data is a valuable asset. However, the level of its value very much depends on the market context. Consequently, it is not straightforward to what extent accumulation of big data creates barriers to entry and market power.

Furthermore, competition advocates have put forward several theories of harm in relation to the development of big data. Personalized pricing and algorithmic collusion are two of them, frequently mentioned. As regards the

[50] See paragraph 350 of the Commission Decision.

[51] See Commission's press release, available at: http://europa.eu/rapid/press-release_IP-16-4284_en.htm

former, it assumes that data and algorithms may lead to a situation of perfect price discrimination where the willingness of pay of consumers is completely revealed. The algorithmic collusion theory of harm is based on the hypothesis that algorithms increase transparency in the market and make tacit coordination between firms much easier than it was in the case of traditional industries.

Competition enforcers are dedicating resources to understand these theories of harm and more generally to assess whether the shift to digitalization of the economy implies any updates in their current work. This chapter has described a few recent decisions by the European Commission where data issues played a role to a certain, albeit limited, extent. Finally, competition policy is not the only field where big data and algorithms have raised waves. Experts in areas such as consumer protection, privacy or ethics are also vigilant and participate vividly in the debate about the pros and cons of digitalization.

BIBLIOGRAPHY

Autorite de la Concurrences and Bundeskartellamt (2016), *Competition Law and Data,* available at: https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Berichte/Big%20Data%20Papier.pdf?__blob=publicationFile&v=2

Bourreau, M.; de Streel, A., and I. Graef (2017), *Big Data and Competition Policy: Market power, personalised pricing and advertising,* Centre on Regulation in Europe, available at: http://www.cerre.eu/sites/cerre/files/170216_CERRE_CompData_FinalReport.pdf

Colangelo, G., and M. Maggiolino (2017), "Big Data as Misleading Facilities," *European Competition Journal,* Bocconi Legal Studies Research Paper No. 2978465, Forthcoming.

Cole, M. (2018), Data in EU Merger Control, *Competition Policy International, Antitrust Chronicle,* February.

Evans, D. S., and R. Schmalensee (2018), "Debunking the 'Network Effects' Bogeyman," *Regulation,* Vol. 40, No. 4, Winter 2017-2018.

Ezrachi, A., and M. E. Stucke (2016), *Virtual competition,* Harvard University Press.

Fumagalli, C.; Motta, M., and C. Calgano (2018), *Exclusionary Practices – The Economics of Monopolisation and Abuse of Dominance,* Cambridge University Press.

FTC (2016), "Big Data: A Tool for Inclusion or Exclusion? – Understanding the Issue." Available at: https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

Graef, I. (2015), "Market Definition and Market Power in Data: The Case of Online Platforms," *World Competition: Law and Economics Review,* Vol. 38, No. 4: 473-506. Available at SSRN: https://ssrn.com/abstract=2657732 or http://dx.doi.org/10.2139/ssrn.2657732

Hansen, S. (2018), "Machine Learning for Economics and Policy," in *Economic Analysis of the Digital Revolution,* Funcas.

Kennedy, J. (2017), *The Myth of Data Monopoly: Why Antitrust Concerns about Data are Overblown.* Available at: https://itif.org/publications/2017/03/06/myth-data-monopoly-why-antitrust-concerns-about-data-are-overblown

Lambrecht, A., and C. E. Tucker (2015), "Can Big Data Protect a Firm from Competition?," *Competition Policy International,* Antitrust Chronicle, January 2017.

Marr, B. (2016), *Big data in practice,* Wiley Publisher.

Mayer-Schonberger, V., and K. Cukier (2013), *Big data – a revolution that will transform how we live, work and think,* Eamon Dolan/Houghton Mifflin Harcourt.

Montes, R.; Sand-Zantman, W., and T. Valletti (2017), *The value of personal information in online markets with endogenous privacy.* Available at: https://www.orange.com/fr/content/download/45273/1347122/version/2/file/Privacy_online_market_October2017.pdf

OECD (2013a), "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value," *OECD Digital Economy Papers,* No. 220, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k486qtxldmq-en

— (2013b), "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by 'big data'," in *Supporting Investment in Knowledge Capital, Growth and Innovation,* OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-12-en

— (2017), *Algorithms and collusion.* Available at: http://www.oecd.org/competition/algorithms-and-collusion.htm

Prüfer, J., and C. Schottmuller (2017), Competing with Big Data, *Discussion Paper,* Vol. 2017-007, Center for Economic Research, Tilburg. Available at: https://schottmueller.github.io/papers/tipping/Competing%20with%20Big%20Data.pdf

Rubinfeld, D. L., and M. S. Gal (2017), "Access Barriers to Big Data (August 26, 2016)," *59 Arizona Law Review,* 339. Available at SSRN: https://ssrn.com/abstract=2830586 or http://dx.doi.org/10.2139/ssrn.2830586

Schneier (2010), *A blog covering security and security technology.* Available at: http://www.schneier.com/blog/archives/2009/11/a_taxonomy_of_s.html

Shapiro, C. (2016), *Did Arrow Hit the Bull's Eye?.* Available at: http://faculty.haas.berkeley.edu/shapiro/arrow.pdf

Stephens-Davidowitz, S. (2017), *Everybody lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are,* HarperCollins Publishers.

Stucke, M. E., and A. P. Grunes (2016), *Big data and competition policy,* Oxford University Press.

About the autors

Luis Aguiar

Is an economist and a Research fellow at the European Commission's Joint Research Center in Seville, Spain. His main research interests are in empirical industrial organization and in the economics of digitization, with a particular focus on the effects of technological change on firms, consumers, markets, and welfare. His research has evaluated the welfare benefits resulting from the introduction of new products and the effects of technological change on the supply of recorded music. He has also conducted empirical studies on the effectiveness of online copyright enforcement policies in the movie industry, the interaction between various music consumption modes and their implications, as well as the effects of music and movie trade. Luis Aguiar holds a PhD in Economics from Universidad Carlos III de Madrid, an MSc in Economics, Finance and Management from Universitat Pompeu Fabra in Barcelona, and a Degree in Economics from the University of Geneva.

Paul Belleflamme

Is Professor of Economics at Université Catholique de Louvain, where he is attached to the Center for Operations Research and Econometrics (CORE) and to the Louvain School of Management (LSM). Paul's main research area is theoretical industrial organization, with a special focus on innovation in the digital economy (which is also the main topic of his blog, www.IPdigIT.eu). Paul has published widely in leading economics journals and is, with Martin Peitz, the author of *Industrial Organization: Markets and Strategies* (Cambridge University Press, 2010 and 2015). Paul is a Fellow of the CESifo Research Network. He is Associate Editor of *Journal of Economics,* and co-editor of Economics *E-Journal* and *Regards Economiques.* He also served as Associate Editor of *Information Economics and Policy,* and of *Review of Network Economics.*

About the autors

Carlos Bellón

Holds undergraduate degrees in both Law and Business Administration from Universidad Pontificia Comillas (ICADE). After four years working in the City of London as an investment banker (Citigroup and JP Morgan) he completed his studies at Wharton (University of Pennsylvania), where he obtained an MBA (2005) and a PhD in Finance (2012). He worked as Assistant Professor in Universidad Carlos III de Madrid for five years and joined the faculty of Universidad Pontificia Comillas in 2017. His research is focused on the finance of innovation and the effect new technologies have on traditional financial institutions (crowdfunding, FinTech, cryptocurrencies, etc.).He is currently also venture partner in a venture capital fund that invests in early financing rounds of FinTech startups.

Adina Claici

Is an expert in competition economics. She is Director of the Brussels Office of Copenhagen Economics where she manages cases in all areas of competition enforcement. Prior to joining the private practice, Adina had been a Senior Member of the Chief Economist Team at DG Competition for almost ten years. She gathered a unique insight in the most recent competition policy developments and in-depth experience in mergers, anti-trust and state aid cases covering many industries. Adina has also significant experience with the interaction between competition policy and regulation across many sectors, including the digital economy. Adina is a regular speaker at conferences and she is publishing in peer-reviewed academic journals. She is currently also a visiting professor at the College of Europe in Bruges and at Barcelona Graduate School of Economics. Adina has a PhD in Economics from Autonoma University of Barcelona.

About the autors

Francesco Decarolis

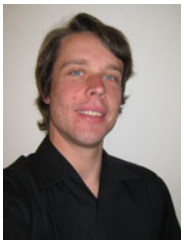
Born May 17, 1980. Degree in Economics cum laude from Università Bocconi in 2002. PhD in Economics from University of Chicago in 2000. Associate Professor, Università Bocconi since 1st November 2017. Full Professor qualification obtained in 2016 (13-A1, A2 and A3 bando D.D. 1532/2016). Associate Professor (with Tenure) at EIEF-Einaudi Institute for Economics and Finance (2016). Assistant Professor at Boston University (2012-2016). Since 2017 European Commission – DG Comp, Economic Advisory Group on Competition Policy (EAGCP). Since 2016 NBER, Faculty Research Fellow and since 2015 CEPR, Research Affiliate, UK. Research areas: Industrial Organization, Applied Microeconomics, Market Design, Health Economics.

Juan-José Ganuza

Is Professor at the department of Economics and Business of the Universitat Pompeu Fabra and affiliated professor of the Barcelona GSE. Juan-José Ganuza holds a bachelor's degree in Physics by the Universidad Complutense de Madrid (1991) and a Ph. D. in Economics by University Carlos III de Madrid (1996). In 1997-98 he was a post doc at California Los Angeles University and Institut D'Economie Industrielle (Toulouse). In 1998 he joined the Universitat Pompeu Fabra, where he was promoted to Full Professor in 2009. His main research interests are: i) auctions, procurement and mechanism design, ii) law and economics, iii) regulation and competition policy and iv) business strategy and innovation. He has published in the main international economics journals in his research field *(RAND Journal of Economics, International Journal of Industrial Organization, Journal of Industrial Economics, Journal of Economics Management and Strategy,* etc.) as well as in general interest economics journals (such as *Econométrica*), law journals (such as the *Journal of Legal Studies*), and business publications (such as *Management Science*). Finally, he has collaborated on several books related with procurement and regulatory issues, among them, *The Handbook of Procurement* (Cambridge University Press, 2006).

About the autors

Maris Goldmanis



Is Senior Lecturer in Economics at Royal Holloway University in London (UK) since 2010. He obtained his PhD in Economics from the University of Chicago. His research –which has been published on leading academic journals such as the *Economic Journal, Managament Science,* and others– focuses on the areas microeconomic theory, industrial organization, and natural resources economics.

Guillaume Haeringer

Specializing in market design, game theory, and social choice, Guillaume Haeringer focuses his research on the design and analysis of market institutions, with a special interest in situations to which the traditional competitive approach is not well suited. Much of his recent investigation concerns the study of matching markets, which involves the use of theoretical advancements to analyze real-life situations and design practical solutions that can be implemented by policymakers. His research on the analysis of mechanisms to assign students to schools and matching markets has been published in the *American Economic Review,* the *Journal of Economic Theory,* and the *International Journal of Game Theory.* Guillaume Haeringer has a bachelor's degree in mathematics and economics and a PhD in economics from the Université de Strasbourg (France). Before being a faculty at Baruch College in New York he held positions at Warwick University and the Autonomous University of Barcelona.

Hanna Halaburda

Is a Visiting Professor at New York University's Stern School of Business and a Senior Economist at the Bank of Canada. Before joining Stern, she was an Assistant Professor at the Harvard Business School. In her research, Hanna studies how technology influences network effects and interactions in the marketplace, and how technology affects business models. Much of her work focuses on competition between platforms, *e.g.* Apple's iPhone vs. Android or eHarmony vs. Match. Another theme in her research is the development of digital

currencies and blockchain technologies. Hanna's work has been published in *Management Science, American Economic Journal,* and *Games and Economic Behavior* and other academic journals. She also wrote a book (joint with Miklos Sarvary) on digital currencies, *Beyond Bitcoin: The Economics of Digital Currency* (Palgrave, 2016). Hanna holds master's degrees in economics (Warsaw School of Economics) and philosophy (Warsaw University), and a PhD in economics (Northwestern University).

Stephen Hansen

Austin (Texas) October, 1981. Is Professor of Economics at University of Oxford. Stephen obtained his Ph.D in Economics from London School of Economics in 2009. He previously worked at Universitat Pompeu Fabra in Barcelona as an Assistant Professor and then as an Associate Professor. He also works in economic consultancy to the Bank of Englad and CEPR Affiliate and Oxford Man Institute. Is a Fellow of the Alan Turing Institute and Associate. He has his work in lending economic journals: *Quarterly Journal of Economic, The review of Economic Studies, Journal of Applied Econometrics, Journal of International Economics, International of Law and Economics, Journal of Monetary Economics* and *Journal of Law Economics, and Organization.*

Doh-Shin Jeon

Is Professor of Economics at Toulouse School of Economics. Doh-Shin is also a part-time Research Fellow at the CEPR. Doh-Shin obtained his Ph.D. in Economics from University of Toulouse 1 in 2000. He previously worked at Universitat Pompeu Fabra in Barcelona as an Assistant Professor and then as an Associate Professor. Expert in industrial organization, Doh-Shin has mainly conducted research on the economics of information technology and intellectual property with particular focus on digital platforms. He published numerous articles in major international journals such as *American Economic Review, Rand Journal of Economics, American Economic Journal: Microeconomics, Journal of the European Economic Association,* etc. He is an Associate Editor of the *Journal of Industrial Economics.*

About the autors

Gerard Llobet

Is associate professor with tenure at CEMFI. He obtained his undergraduate degree in Economics and Management from Universitat Pompeu and his PhD in Economics from the University of Rochester in 2000. His research spawns areas in industrial economics and innovation. His recent works study how intellectual property rights affects the incentives for firms to innovate and contribute their technology in Standard Setting Organizations. He has published in outlets like the *Journal of Political Economy, Management Science, Review of Financial Studies or the Journal of Law and Economics.* He is also active in making economic results available to general audiences. He is a regular contributor to the blog Nada es Gratis of which he has also been an editor.

Jose Luis Moraga

Is currently professor of Microeconomics at the Vrije Universiteit Amsterdam, and professor of Industrial Organization at the University of Groningen. He obtained a Ph.D. in Economics from University Carlos III Madrid in 1997 and became a post-doctoral researcher at the Institute of Economics in Copenhagen. From 1998 to 2005 he worked at Erasmus University Rotterdam and Tinbergen Institute. In 2005 he moved to the University of Groningen where he was Research Director of the Institute of Economics, Econometrics and Finance and team leader of a Marie Curie Excellence Grant on search and switching costs. José L. main research interests pertain to the imperfections of markets arising from price-setting power. He has conducted research on various factors contributing to market power such as transaction costs, advertising costs, forward selling, research and development, etc. José L. has applied insights from industrial organization to energy economics and international economics, specifically to the analysis of forward contracting and antidumping legislation. He has published his work in leading economic journals, including the *Journal of Political Economy, The Economic Journal, The Review of Economic Studies and the Rand Journal of Economics.* He is currently co-editor of the *International Journal of Industrial Organization* and associate editor of the *Journal of Industrial Economics.*

About the autors

Martin Peitz

Is Professor of Economics at the University of Mannheim since 2007 and a Director of the Mannheim Centre for Competition and Innovation (MaCCI) since 2009. Martin Peitz has been Editor of *Review of Network Economics* and co-editor of *International Journal of Industrial Organization.* He is associate editor of *Journal of Industrial Economics* and *Information Economics and Policy.* He is Research Fellow of CEPR, CESifo, and CERRE; CRESSE Associate; and ZEW Research Associate. He has widely published in leading economics journals. He is author of the books *Industrial Organization: Markets and Strategies* (with Paul Belleflamme) and *Regulation and Entry into Telecommunications Markets* (with Paul de Bijl), both published by Cambridge University Press, and editor of the *Oxford Handbook of the Digital Economy* (with Joel Waldfogel), published by Oxford University Press. His research covers various topics in industrial organization, competition policy, regulation, and microeconomics.

Antonio Penta

Is Professor of Economics at the Universitat Pompeu Fabra and the Barcelona GSE since September 2017. Prior to that, he was Assistant Professor and then tenured Associate Professor in Economics at the University of Wisconsin-Madison. He graduated from Universita' Bocconi (Milan, Italy) in 2004, and obtained a M.A. (2008) and a PhD in Economics (2010) from the University of Pennsylvania (Philadelphia, USA). He is currently an Associate Editor of the *Journal of Economic Theory.* In 2017 he was awarded a European Research Council (ERC) Starting Grant, which among other themes encompasses research on Online Auctions and Digital Marketing Agencies. His research –which has been published in leading academic journals such as the *American Economic Review, Econometrica, Journal of Economic Theory, Review of Economic Studies,* and others– spans several areas of economic theory, such as game theory, mechanism design, bounded rationality and auction theory.

About the autors

Michelangelo Rossi

Is a current PhD student at the Department of Economics at the Universidad Carlos III de Madrid (UC3M). His research interests center around the role of *review systems* in digital platforms. Specifically, his works examine the capacity of reviews to *signal quality* and to *monitor users' behavior.* His study deals with reviews in forms of *numerical ratings* and *textual comments,* whose content is explored using sentimental analysis techniques. In his current work, he uses a textual analysis over Airbnb guests' comments to estimate the dynamics of effort that hosts exert in each transaction over their life-cycle. He works under the supervision of two professors: Natalia Fabra (UC3M) and Matilde Machado (UC3M).

Pablo Ruiz-Verdú

Is an Associate Professor of Management at the Department of Business Administration at Universidad Carlos III de Madrid. He holds a Ph. D. in Economics from Stanford University and a Bachelor's degree in Economics from Universidad Carlos III de Madrid. He teaches Economics of Organizations and Business Economics to undergraduates and doctoral students. Pablo's research interests center on corporate governance, the behavior and impact of institutional investors, human resource economics, and the financing of innovation. His work has been published in journals such as *The Journal of Finance, Review of Finance, The Journal of Economic Behavior and Organization, European Financial Management, Labour Economics or Economics Letters.*

Juan Manuel Sánchez-Cartas

Graduated in Business Administration (2011) and Economics (2013) at the University of Extremadura. He also holds a Master of Science in Applied Economics from the Alcala University and the Complutense University of Madrid (2015). Currently, he is a Ph.D. Candidate at the Technical University of Madrid and at the European Institute of Technology. He has been a Research Assistant at the Complutense University

of Madrid and a Visiting Researcher at the University of Zürich. He has also been an advisor of Red.es and Telefónica Spain on digital ecosystems and platforms. Juan Manuel's main research area is the theoretical industrial organization, with a special focus on multi-sided markets and compatibility issues. He also works on other areas, such as digital markets, market simulations, and digital ecosystems.

Mateo Silos

Is Principal Economist at Ofwat, the economic regulator and competition authority for the water sector in England and Wales. He has spent most of his professional career working in the area of competition policy and regulation. Before joining Ofwat in November 2017, he was Head of Economic Analysis at the Spanish Competition and Markets Commission. Previously, he worked as an economist at the former Spanish Competition Commission. He has also worked in economic consultancy, both in London and Madrid, an at the Lawrence R. Klein Institute for Economic Forecasting (Madrid). He holds a Master's degree in Economics from Universidad Complutense de Madrid and Bachelor's degrees in Economics and Philosophy from Universidad Autónoma de Madrid.

Joel Waldfogel

Is the Frederick R. Kappel Chair in Applied Economics at the University of Minnesota's Carlson School of Management. Before coming to Minnesota, Waldfogel was at the University of Pennsylvania's Wharton School (1997-2010), the Yale University Economics Department (1990-1997). Waldfogel has conducted empirical studies of price advertising, media markets, the operation of differentiated product markets, and issues related to digital products, including piracy, pricing, revenue sharing, and the effects of digitization on the supply of new products. Most of his research since 2004 has focused on copyright-related issues. He has published papers on the impact of piracy on the revenue of authorized products in recorded music, motion pictures, and television. Since 2010 he

has done research, and has published papers, on the broader impacts of digitization on the supply of new products in creative industries. His book *Digital Renaissance* (Princeton University Press) appears in fall 2018.

Published issues

Nº 1. Monetary Policy after the Great Recession.
Edited by Javier Vallés

Nº 2. The *Triple Aim* for the future of health care.
Edited by Núria Mas and Wendy Wisbaum

Nº 3. The voice of society in the face of the crisis.
Edited by Víctor Pérez-Díaz

Nº 4. Financial History Workshop. Improving Savings Culture.
A Lifetime of Financial Education

ECONOMIC ANALYSIS OF THE DIGITAL REVOLUTION

"Instead of trying to produce a program to simulate the adult mind, why not try to produce one which simulates the child's mind? If it is then submitted to an appropriate educational course adult's brain would get".

A. M. Turing (1950)
Computing Machinery and Intelligence. **Mind 49: 433-460**

Funcas

Caballero de Gracia, 28
28013 Madrid
Teléfono: 91 596 54 81
Fax: 91 596 57 96

publica@funcas.es
www.funcas.es

Electronic version available at:
http://www.funcas.es/Publicaciones

ISBN: 978-84-17609-00-9